МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ АГЕНТСТВО ПО ОБРАЗОВАНИЮ

ГОУ ВПО «СИБИРСКАЯ ГОСУДАРСТВЕННАЯ ГЕОДЕЗИЧЕСКАЯ АКАДЕМИЯ»

Ю.А. Кравченко

ОСНОВЫ КОНСТРУИРОВАНИЯ СИСТЕМ ГЕОМОДЕЛИРОВАНИЯ

Книга 2

ИНФОРМАЦИОННОЕ ГЕОМОДЕЛИРОВАНИЕ: МОДЕЛИ И МЕТОДЫ

Часть 1

Новосибирск
СГГА
2008

УДК 528.91
К772

Рецензенты:

Доктор технических наук, профессор
Томского государственного университета
А.В. Скворцов

Кандидат технических наук, доцент Новосибирского
государственного архитектурно-строительного университета
А.Ф. Задорожный

Кравченко, Ю.А.
К772 Основы конструирования систем геомоделирования. Книга 2. Информационное геомоделирование: модели и методы. Часть 1 [Текст] : монография / Ю.А. Кравченко. – Новосибирск: СГГА, 2008. – 315 с.

ISBN 978-5-87693-303-4 (ч. 1)
ISBN 978-5-87693-302-7 (кн. 2)
ISBN 978-5-87693-296-9

Рассматривается проблема представления геопространства в целом в системах информационного геомоделирования, дается решение главных геодезических задач на поверхности эллипсоида вращения. Описываются структура моделей топографических поверхностей и методы их создания.

Для студентов старших курсов, аспирантов и специалистов в области информационного геомоделирования, геоинформатики и картографии.

Печатается по решению редакционно-издательского совета СГГА

Научный редактор: кандидат технических наук,
профессор Сибирской государственной геодезической академии
Ю.Г. Костына

УДК 528.91

ISBN 978-5-87693-298-3 (ч. 1)
ISBN 978-5-87693-302-7 (кн. 2)
ISBN 978-5-87693-296-9

СОДЕРЖАНИЕ

7. ГЕОМЕТРИЧЕСКИЕ ЗАДАЧИ В СИСТЕМАХ ГЕОМОДЕЛИРОВАНИЯ

Системы геомоделирования в конечном итоге предназначены для решения тех или иных прикладных задач. Если рассматривать все множество таких задач, то их можно разделить на три самых крупных блока:

1) специфические задачи, характерные для конкретной проблемной области; примерами могут служить расчет электрических сетей, вычисление объемов работ при открытых способах разработки полезных ископаемых, поиск кратчайшего пути между двумя точками земной поверхности, определение зон радиовидимости при проектировании сетей сотовой связи и т. п.;

2) поисковые задачи, связанные с извлечением семантической информации, нахождением объектов, обладающих заданными свойствами;

3) геометрические задачи на земной поверхности, общие для любых систем геомоделирования: вычисление координат точек, расстояний, углов, площадей и т. п.

Первые из указанных задач не имеет смысла рассматривать в книге, посвященной решению общих проблем геоинформационного моделирования. Вторые задачи решаются с использованием штатных средств систем управления реляционными базами данных, и современные методы их решения рассматривались в одной из предыдущих глав. Наконец, необходимость решения третьих задач является отличительным признаком систем геоинформационного моделирования, выделяющим их из всего множества информационных систем. Таким образом, средства решения геометрических задач на земной поверхности являются важным компонентом систем геомоделирования.

В настоящее время в геоинформационных системах геометрические задачи принято решать преимущественно на плоскости в какой-либо картографической проекции. Эта методология была разработана задолго до появления вычислительных машин и главная роль в ней отводилась человеку. Поэтому наибольшее внимание в ней уделялось получению наиболее простых формул для массовых вычислений. При необходимости человек измерял исходные величины по карте и выполнял несложные вычисления. Теоретической основой указанной методологии являлись *математическая картография* и такой раздел картографии, как *картометрия*. Преимущество указанной методологии состоит в простоте и доступности применяемых формул; обычно достаточно использования средств элементарной или аналитической геометрии на плоскости и тригонометрии. Недостатком решения геопространственных задач на плоскости является неудовлетворительная точность величин, определяемых по карте, поскольку, во-первых, линейные величины по карте или плану не могут быть измерены точнее, чем 0,1 мм (так называемая *точность карты*), и, во-вторых, любая картографическая проекция характеризуется искажениями геометрических величин, которые тем больше, чем больше изображаемая поверхность.

С тех времен, когда была разработана традиционная методология, требования к точности решения задач на земной поверхности многократно повысились, а

точность определения координат на ней с помощью инструментальных средств (*GPS*) возросла в десятки раз. Однако способы решения геометрических задач остались прежними и благополучно перекочевали в ГИС.

Причина такого состояния дел, видимо, заключается в следующем. Специалистам, имеющим геодезическое или картографическое образование и работающим в геоинформатике, как и любым другим людям, присуща определенная степень консерватизма. Они действуют так, как их в свое время обучали в вузах. С другой стороны, в практическую геоинформатику пришло большое количество людей, не имеющих геодезического или картографического образования и вообще какого-либо отношения к наукам о Земле. В публикациях можно найти сведения о прецедентах реализации геоинформационных проектов даже психологами. Естественно, эти люди ничего не подозревают о несовершенстве используемой методологии.

Но не все столь безнадежно, как это может показаться. На этом фоне выделяются специалисты, получившие физико-математическое образование. И можно назвать российские коллективы, добившиеся серьезных результатов в создании систем геомоделирования. Однако, их разработки в области решения геометрических задач на земной поверхности также в основном следуют общепринятой методологии.

Здесь можно было бы перейти к более детальному рассмотрению недостатков использования традиционных методов решения геометрических задач в системах геомоделирования. Но, поскольку данную книгу будут читать не только геодезисты, постольку необходимо сделать некоторый экскурс в сфероидическую геодезию, в противном случае содержание данной главы будет непонятным более широкому кругу читателей.

Также возможно, что данная глава будет небесполезной для специалистов по кадастровым системам. Автору несколько лет назад пришлось знакомиться с вкладышем в диплом одного молодого специалиста по кадастру. Как оказалось, он изучал четыре вида права и такой замечательный предмет, как «проектирование карьеры», но не имел никакого представления о сфероидической геодезии.

Сфероидическая геодезия входит в курс высшей геодезии, который делится на две части: сфероидическую геодезию и физическую геодезию. Сфероидическая геодезия – один из краеугольных камней геодезического знания – является приложением геометрии к решению практических задач на эллипсоиде. При изучении геометрии вообще большое значение имеет наглядность, однако изложение сфероидической геодезии в учебной литературе, по мнению автора, носит недостаточно последовательный характер, что несколько затрудняет ее понимание и изучение.

В наши намерения не входит полное переписывание учебников по сфероидической геодезии, хотя необходимость в этом существует. Как представляется, излагаемый ниже материал может служить достаточно ценным дополнением к ним. Для более глубокого ознакомления с предметом сфероидической геодезии можно рекомендовать [4] и [7].

7.1. Эллипс

Основными математическими объектами, изучаемыми в сфероидической геодезии, являются эллипс и фигура, получающаяся при его вращении вокруг малой оси, – эллипсоид вращения. Таким образом, эллипс, в известном смысле, первичен, поскольку определяет форму эллипсоида. Следовательно, изучение эллипса имеет принципиальное значение.

Эллипс – замкнутая центральная линия второго порядка, симметричная относительно осей *Ox* и *Oy* – определяется как кривая, все точки которой обладают тем свойством, что сумма расстояний от каждой из них до двух заданных точек, называемых *фокусами*, есть величина постоянная:

$$r_1 + r_2 = const = 2a\,. \qquad (7.1)$$

Из этого условия можно получить *каноническое уравнение эллипса*:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1, \qquad (7.2)$$

где a и b – его *большая* и *малая полуоси* соответственно.

Эллипс представлен на рис. 7.1, где точки F_1 и F_2 – фокусы эллипса, а прямые d_1 и d_2 – так называемые *директрисы*. Середина расстояния между фокусами является *центром эллипса*.

Размеры эллипса определяются его малой или большой полуосью, а форма (вытянутость) – величиной, называемой *эксцентриситетом*. В математической литературе эксцентриситет e эллипса обычно определяется как «отношение расстояния любой точки эллипса до фокуса к расстоянию ее до соответствующей директрисы» [6, с. 649]:

$$e = \frac{r_1}{d_1} = \frac{r_2}{d_2}. \qquad (7.3)$$

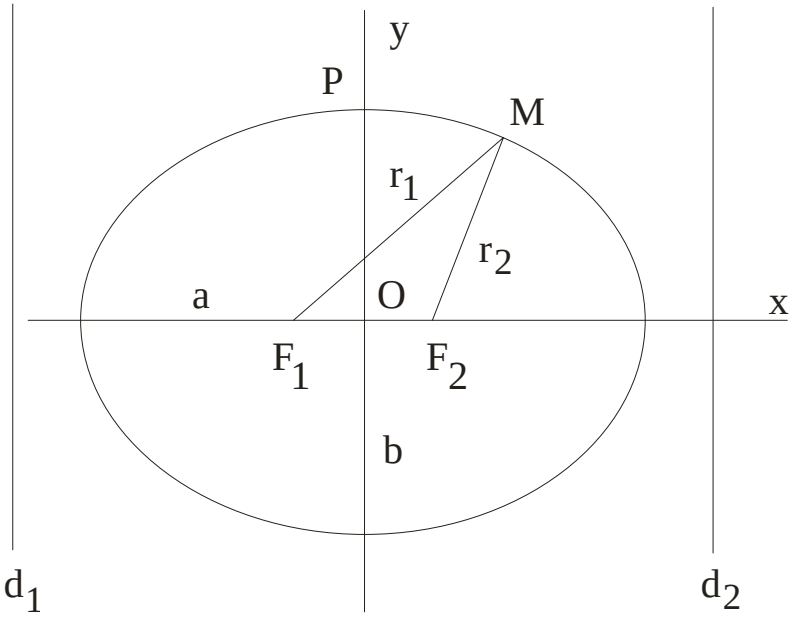


Рис. 7.1. Эллипс

В геодезической литературе эксцентриситет эллипса вводят как соотношение между его малой и большой полуосями:

$$e = \frac{\sqrt{a^2 - b^2}}{a} \qquad (7.4)$$

и называют также *первым эксцентриситетом*. Кроме того, иногда используется *второй эксцентриситет*, определяемый выражением

$$e' = \frac{\sqrt{a^2 - b^2}}{b}, \qquad (7.5)$$

и сжатие эллипса

$$\alpha = \frac{a - b}{a}. \qquad (7.6)$$

При $e = 0$ эллипс превращается в окружность. Расстояние от центра эллипса до фокуса составляет

$$x = \pm ae, \qquad (7.7)$$

от центра эллипса до директрисы

$$x = \pm \frac{a}{e}, \qquad (7.8)$$

а от фокуса до директрисы равно

$$d = \pm \frac{p}{e},$$

где p – *фокальный параметр*, равный половине хорды, проходящей через фокус и параллельной малой оси

$$p = \frac{b^2}{a}; \qquad (7.9)$$

следовательно, можно найти, что

$$d = \frac{b}{e'}. \qquad (7.10)$$

Как геометрическая фигура, эллипс обладает такими замечательными свойствами:

1) сумма расстояний от произвольной точки эллипса до его фокусов есть величина постоянная, равная $2a$, что следует непосредственно из его определения;

2) нормаль к эллипсу в произвольной точке M делит пополам угол F_1MF_2 (см. рис. 7.1) между прямыми, соединяющими эту точку с фокусами (так называемое *фокальное свойство эллипса*).

Расстояние от точки полюса P до любого фокуса равно a, что следует из первого указанного свойства эллипса.

Значение *полярного радиуса c* есть величина

$$c = \frac{a^2}{b} = \frac{a}{\sqrt{1 - e^2}}. \qquad (7.11)$$

Значения не получивших общепринятых названий, но иногда используемых величин n и m задаются выражениями:

$$n = \frac{a-b}{a+b}; \qquad (7.12)$$

$$m = \frac{a^2 - b^2}{a^2 + b^2}. \qquad (7.13)$$

Радиус кривизны эллипса в точке M определяется выражением

$$R = a^2 b^2 \left(\frac{x^2}{a^4} + \frac{y^2}{b^4} \right)^{3/2}. \qquad (7.14)$$

Положение произвольной точки M на эллипсе определяется значением ее широты (рис. 7.2). При этом в геодезии принято различать *геоцентрическую широту* φ точки – угол между ее радиус-вектором и осью абсцисс и *геодезическую широту* B – угол между нормалью к эллипсу в этой точке и осью абсцисс. Наиболее простое соотношение между геодезической и геоцентрической широтой определяется выражением

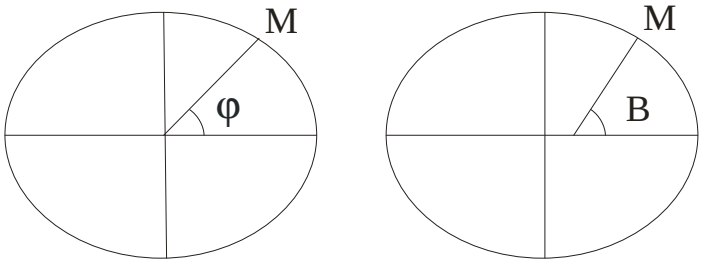


Рис. 7.2. Широты точки на эллипсе

$$\operatorname{tg}\varphi = (1 - e^2)\operatorname{tg}B, \qquad (7.15)$$

из которого можно получить другие формулы. Кроме того, очень часто используется значение *приведенной широты*, которую принято обозначать как u, и интерпретация которой будет дана дальше. Зависимость между приведенной и геоцентрической широтами точки выражается формулой

$$\operatorname{tg}\varphi = \sqrt{1 - e^2}\operatorname{tg}u. \qquad (7.16)$$

В полярной системе координат, полюс которой совпадает с фокусом F_2, а полярная ось направлена по оси Ox, уравнение эллипса принимает вид

$$\rho = \frac{p}{1 + \cos\theta}, \qquad (7.17)$$

где θ – угол между полярной осью и радиус-вектором точки на эллипсе; величина p – фокальный параметр.

Площадь эллипса равна $S = \pi ab$.

7.2. Альтернативная интерпретация параметров эллипса

Ниже дается другая трактовка эксцентриситетов эллипса. Известно, что эллипс может быть получен сечением прямого кругового конуса или цилиндра плоскостью. На рис. 7.3 представлены эллипс, полученный сечением прямого кругового цилиндра плоскостью под некоторым углом $90^\circ - \varepsilon$ к оси цилиндра, и

окружность, образовавшаяся при сечении этого же цилиндра плоскостью, перпендикулярной к его оси. Линия P_1P_2 является линией пересечения плоскости эллипса с плоскостью окружности; точка *O* является центром и эллипса, и окружности. Радиус окружности обозначим *b*. В плоскости окружности введем систему координат с центром в точке *O* таким образом, чтобы ось *Ox* была перпендикулярна линии P_1P_2 , а ось *Oy* совпадала с последней. В плоскости эллипса введем систему прямоугольных координат *OXY*, ось *OX* которой перпендикулярна линии P_1P_2, а ось *OY* совпадает с линией P_1P_2 . Угол ε между осями координат *Ox* и *OX* равен двугранному углу между плоскостью эллипса и плоскостью окружности.

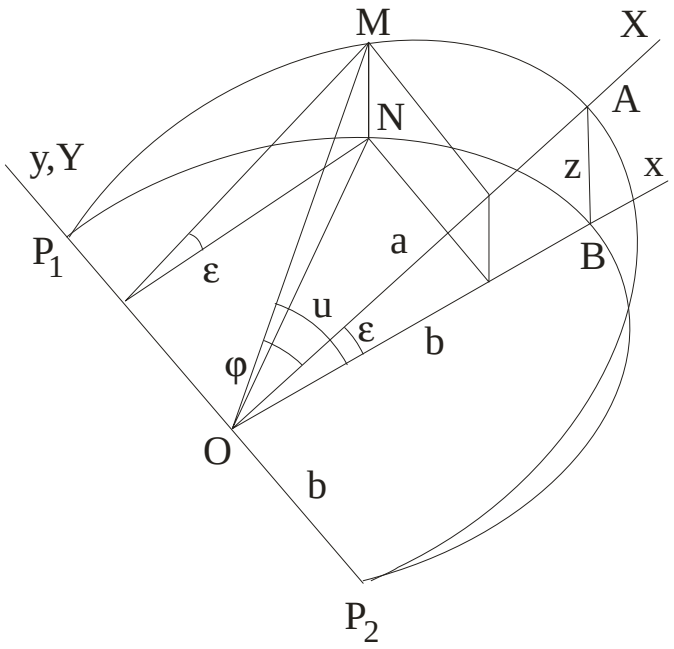


Рис. 7.3. К определению эксцентриситета

Из рис. 7.3 следует, что отрезок прямой *OA* является большой полуосью *a* эллипса, а радиус окружности *b* одновременно является его малой полуосью. Нетрудно видеть, что значение первого эксцентриситета равно синусу угла между плоскостью окружности и плоскостью эллипса:

$$e = \frac{\sqrt{a^2 - b^2}}{a} = \sin\varepsilon, \qquad (7.18)$$

значение второго эксцентриситета – это тангенс угла ε:

$$e' = \frac{\sqrt{a^2 - b^2}}{b} = \mathrm{tg}\varepsilon, \qquad (7.19)$$

а отношение полуосей – его косинус:

$$\frac{b}{a} = \cos\varepsilon. \qquad (7.20)$$

При этом сжатие эллипса α есть дополнение $\cos\varepsilon$ до 1:

$$\alpha = \frac{a-b}{a} = 1 - \cos\varepsilon. \qquad (7.21)$$

Значение полярного радиуса *c* определяется выражением

$$c = \frac{a^2}{b} = \frac{a}{\sqrt{1-e^2}} = \frac{a}{\cos\varepsilon}. \qquad (7.22)$$

Часто используемая в сфероидической геодезии величина n есть

$$n = \frac{a-b}{a+b} = \operatorname{tg}^2 \frac{\varepsilon}{2}. \qquad (7.23)$$

Между n и $\cos \varepsilon$, а также между вспомогательной величиной m и $\cos$ ε можно заметить определенную симметрию:

$$n = \frac{1-\cos\varepsilon}{1+\cos\varepsilon}; \qquad (7.24)$$

$$\cos\varepsilon = \frac{1-n}{1+n}; \qquad (7.25)$$

и

$$m = \frac{1-\cos^2\varepsilon}{1+\cos^2\varepsilon}; \qquad (7.26)$$

$$\cos^2\varepsilon = \frac{1-m}{1+m}. \qquad (7.27)$$

Встречающееся в формулах сочетание $1 + n$ также просто выражается через ε:

$$1+n = \frac{1}{\cos^2 \frac{\varepsilon}{2}}. \qquad (7.28)$$

Предложенная интерпретация первого и второго эксцентриситетов делает некоторые формулы сфероидической геодезии очевидными и почти ненужными. Так, если формула

$$b = a\sqrt{1-e^2}$$

не очевидна и требует некоторых размышлений, то та же зависимость, но выраженная с учетом (7.20) как

$$b = a\sqrt{1-\sin^2\varepsilon},$$

представляется уже банальностью.

Аналогичные утверждения справедливы для ряда других формул сфероидической геодезии. Можно сравнить, например, два выражения, имеющие один и тот же смысл, но разную форму записи

$$e'^2 = \frac{e^2}{1-e^2}$$

и

$$\operatorname{tg}^2\varepsilon = \frac{\sin^2\varepsilon}{1-\sin^2\varepsilon}.$$

Далее мы можем использовать e как краткое обозначение выражения $\sin\varepsilon$.

Координаты точки *M*, принадлежащей эллипсу, как можно установить из рис. 7.3, выражаются через x и y формулой

$$\left.\begin{aligned} X_M &= \frac{1}{\cos\varepsilon}\,x \\ Y_M &= y \end{aligned}\right\}, \qquad (7.29)$$

ее радиус-вектор равен

$$\rho_M = \sqrt{X_M^2 + Y_M^2} = \sqrt{\frac{1}{\cos^2\varepsilon}\,x^2 + y^2}, \qquad (7.30)$$

а угол между ним и осью *OX* определяется выражением

$$\operatorname{tg}\varphi_M = \frac{Y_M}{X_M} = \cos\varepsilon\,\frac{y}{x}. \qquad (7.31)$$

Из исходной окружности с радиусом b можно получить еще один эллипс с тем же значением эксцентриситета $\sin\varepsilon$, но уже как ее ортогональную проекцию на плоскость эллипса (рис. 7.4). Полученную таким образом пару эллипсов можно называть *сопряженными*, при этом эллипс на рис. 7.3 будем называть *большим*, а на рис. 7.4 – *малым*. Малый эллипс развернут относительно большого на 90°, и его большая полуось равна b, а малая полуось – $b\cos\varepsilon$, что следует непосредственно из рис. 7.4.

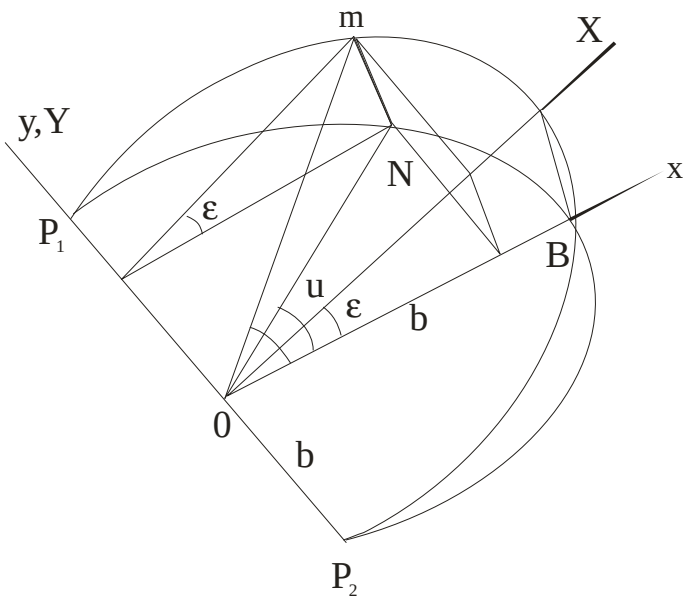


Рис. 7.4. Сопряженный эллипс

Ортогональную проекцию точки *N* на плоскость эллипса обозначим *m*; ее координатами будут

$$\left.\begin{aligned} X_m &= \cos\varepsilon\,x \\ Y_m &= y \end{aligned}\right\}. \qquad (7.32)$$

Величина радиус-вектора точки *m* и его ориентация относительно оси *OX* выражаются соответственно формулами

$$\rho_m = \sqrt{X_m^2 + Y_m^2} = \sqrt{\cos^2\varepsilon\,x^2 + y^2}; \qquad (7.33)$$

$$tg\,\varphi_m = \frac{Y}{X} = \frac{1}{\cos\varepsilon}\,\frac{y}{x}. \qquad (7.34)$$

Если плоскость эллипсов повернуть на угол ε и совместить с плоскостью окружности, то получим рис. 7.5. Так как ординаты точек *M*, *N* и *m* совпадают, то указанные точки лежат на одной прямой, параллельной оси *Ox*. Если точка *N*

движется по окружности, то синхронно с ней перемещаются ее проекции: точки M и m по большому и малому эллипсам соответственно, а также прямая, проходящая через эти точки и параллельная оси абсцисс.

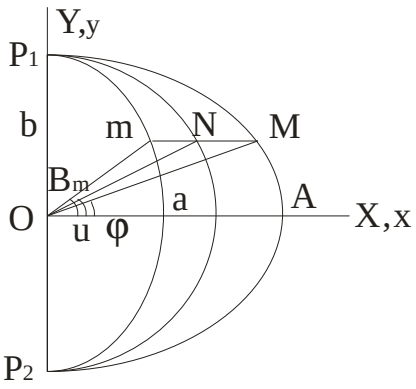


Рис. 7.5. Проекции на плоскость окружности

Из сравнения (7.31) и (7.34) следует

$$\operatorname{tg}\varphi_m = \frac{1}{\cos^2 \varepsilon} \operatorname{tg}\varphi_M . \qquad (7.35)$$

С учетом (7.15) получаем

$$\varphi_m = B_M . \qquad (7.36)$$

В результате мы установили, что значение геоцентрической широты φ точки m малого эллипса, отсчитываемой от оси Ox, равно значению геодезической широты соответствующей точки M большого эллипса. Геоцентрическая широта точки m является ортогональной проекцией приведенной широты u на плоскость эллипса. Но приведенная широта u также суть ортогональная проекция (см. выше), следовательно, геоцентрическая широта φ_m является суперпозицией двух операций ортогонального проектирования.

Предложенная интерпретация первого и второго эксцентриситетов эллипса позволяет получить некоторые зависимости более простым путем. Пусть M – произвольная точка эллипса, а N – ее ортогональная проекция на плоскость окружности (см. рис. 7.3). Тогда расстояние $ON = b$, OM – радиус-вектор ρ точки M; угол φ в плоскости эллипса – *геоцентрическая широта* точки M, а угол u в плоскости окружности – *приведенная широта* точки М. Чтобы показать это, рассмотрим значение тангенса геоцентрической широты:

$$\operatorname{tg}\varphi = \frac{Y}{X} = \frac{y}{\dfrac{x}{\cos\varepsilon}} = \cos\varepsilon \cdot \operatorname{tg}u . \qquad (7.37)$$

Данное выражение есть известная в геодезии зависимость между геоцентрической и приведенной широтами. Таким образом, мы установили простой геометрический смысл приведенной широты: угол u – это ортогональная проекция геоцентрической широты φ на плоскость окружности, которую можно называть *приведенной окружностью*. Чтобы оценить простоту такой интерпретации, можно сравнить, каким образом вводится понятие приведенной широты в учебниках [4] и [7].

Из рис. 7.4 столь же просто установить соотношение между приведенной и геодезической широтой:

$$\operatorname{tg}B = \frac{1}{\cos\varepsilon} \operatorname{tg}u . \qquad (7.38)$$

Используя (7.33), значение радиус-вектора малого эллипса можно представить как функцию приведенной широты

$$\rho_m = b\sqrt{\cos^2\varepsilon\cos^2 u + \sin^2 u}\,. \qquad (7.39)$$

Однако в сфероидической геодезии значение радиус-вектора принято выражать как функцию геодезической широты *B*. Из (7.38) и последнего соотношения можно получить

$$\rho_m = \frac{b\cos\varepsilon}{\sqrt{1-e^2\sin^2 B}}. \qquad (7.40)$$

Выражение в знаменателе называют *первой геодезической величиной* и обозначают как

$$W = \sqrt{1-e^2\sin^2 B}\,. \qquad (7.41)$$

Тогда мы можем (7.40) записать как

$$\rho_m = \frac{b\cos\varepsilon}{W}. \qquad (7.42)$$

Отсюда следует, что первая геодезическая величина *W* является величиной, обратно пропорциональной значению радиус-вектора малого эллипса.

Значение ρ_m можно также представить на основании (7.40) в виде

$$\rho_m = \frac{b}{\sqrt{1+\mathrm{tg}^2\varepsilon\cos^2 B}}, \qquad (7.43)$$

где знаменатель есть так называемая *вторая геодезическая величина*

$$V = \sqrt{1+\mathrm{tg}^2\varepsilon\cos^2 B}\,. \qquad (7.44)$$

С учетом последнего обозначения для вычисления величины радиус-вектора ρ_m можно использовать формулу

$$\rho_m = \frac{b}{V}. \qquad (7.45)$$

Таким образом, вторая геодезическая величина также обратно пропорциональна значению радиус-вектора малого эллипса.

Первую и вторую геодезические величины называют также *основными сфероидическими функциями*. Из сопоставления формул (7.42) и (7.45) следует

$$W = \cos\varepsilon\, V\,. \qquad (7.46)$$

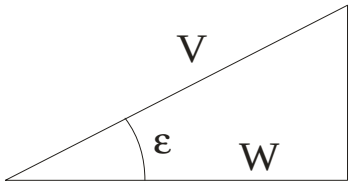


Рис. 7.6. Соотношение между W и V

Поэтому *W* и *V* можно рассматривать как стороны прямоугольного треугольника (рис. 7.6), между которыми существуют соотношения

$$V^2 - W^2 = V^2\sin^2\varepsilon = W^2\mathrm{tg}^2\varepsilon \qquad (7.47)$$

$$V - W = V(1-\cos\varepsilon) = \alpha V; \qquad (7.48)$$

$$V + W = \frac{e^2}{\alpha} V \,. \qquad (7.49)$$

Предложенное истолкование параметров формы эллипса имеет методологическое значение, поскольку облегчает понимание некоторых зависимостей в сфероидической геодезии, а также подсказывает направление преобразований при выводе тех или иных аналитических выражений.

В качестве стандартного значения параметра формы того или иного земного эллипсоида можно указывать значение только угла ε. В частности, для эллипсоида Красовского его значение составляет $\varepsilon = 4°41'34{,}0939037''$.

Очевидно, что обозначения e и e' стали привычными, что не сделало их более понятными. Поэтому представляется целесообразным ввести новые обозначения и интерпретацию параметра формы эллипса и эллипсоида в повседневную геодезическую практику.

7.3. Интерпретация главных геодезических величин

В традиционном изложении курса сфероидической геодезии эллипс рассматривается как некоторая данность, в частности, как функция, описываемая выражением (7.2), из анализа которого выводятся все свойства эллипса. Такой подход является достаточно плодотворным – теория сфероидической геодезии разработана, но все-таки в ней еще существуют некоторые неясности. Примером может служить первая (или вторая) геодезическая величина. Основанием для выделения первой геодезической величины как некоторой сущности служит то обстоятельство, что она часто встречается в формулах сфероидической геодезии [7]. В [4] первая и вторая геодезические величины вводятся вообще без каких-либо комментариев. Хотя выше было дано их геометрическое истолкование, оно мало что прояснило и некоторые вопросы по-прежнему остаются. В частности: «Почему это нечто столь часто встречается в выражениях, что заслужило почетный титул первой геодезической величины?».

Альтернативой традиционному изучению сфероидической геодезии может служить подход, основанный на отображении одного множества (евклидовой плоскости) на другое и дающий возможность ответить на вопросы, подобные сформулированному. Свойства получаемых в результате отображения геометрических объектов будут определяться свойствами исходных фигур и свойствами отображения.

Первая попытка такого подхода к изучению эллипса была предпринята выше, где предложены новая интерпретация и обозначения параметра его формы. Несмотря на безусловную полезность такой трактовки в практическом и теоретическом плане, предлагаемый подход является недостаточно общим. Поэтому наша ближайшая цель – восполнить указанный недостаток и дать более общую точку зрения, когда эллипс рассматривается не как результат сечения кругового цилиндра плоскостью, а как результат проектирования

плоскости с окружностью на другую плоскость, непараллельную первой. Интерпретация первого эксцентриситета как синуса угла между плоскостями и все формулы, приведенные выше, при этом остаются справедливыми. Одновременно появляются новые возможности вывода и объяснения некоторых выражений.

Пусть Q – исходная плоскость, а P – *плоскость проектирования*, или *картинная плоскость* (рис. 7.7). Рассмотрим параллельную проекцию плоскости Q на плоскость P, при которой центр проектирования располагается в бесконечно удаленной точке, а направление проектирования N лежит в плоскости, перпендикулярной линии пересечения плоскостей Q и P.

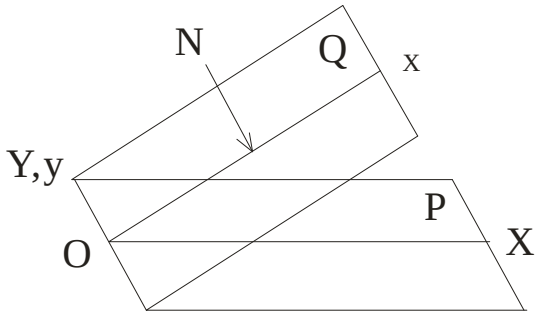


Рис. 7.7. Исходная и проективная плоскости

На плоскости Q введем систему прямоугольных координат Oxy таким образом, чтобы ее ось Ox была перпендикулярна линии пересечения плоскости Q с плоскостью P. На плоскости P введем систему координат OXY, ось OX которой также будет перпендикулярна линии пересечения плоскостей.

Рассмотрим проектирование точки A, принадлежащей плоскости Q, на плоскость P (рис. 7.8). На данном чертеже ε – угол между плоскостями Q и P, ν – угол между нормалью AC к плоскости Q и направлением проектирования AB, отсчитываемый против часовой стрелки.

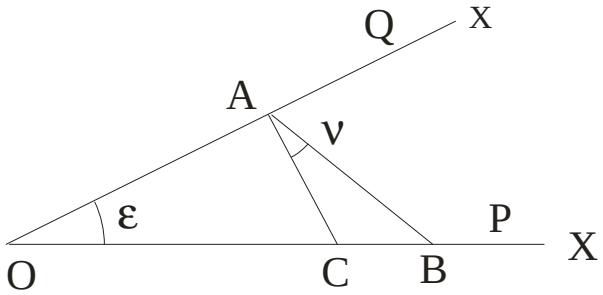


Рис. 7.8. Проекция точки А

Из треугольника OAB получаем соотношение между абсциссой x произвольной точки исходной плоскости и абсциссой X ее проекции на картинную плоскость:

$$X = \frac{\cos\nu}{\cos(\varepsilon+\nu)}x. \qquad (7.50)$$

Обозначив

$$k = \frac{\cos\nu}{\cos(\varepsilon+\nu)}, \qquad (7.51)$$

запишем (7.50) как

$$X = kx. \qquad (7.52)$$

Ордината Y проекции точки равна ординате этой точки:

$$Y = y. \qquad (7.53)$$

Зависимость между приращениями координат выражается также с помощью простых равенств:

$$\left.\begin{aligned}\Delta X &= k\,\Delta x\\ \Delta Y &= \Delta y\end{aligned}\right\}. \qquad (7.54)$$

Выражения (7.54) дают возможность получить зависимость между ориентацией отрезка прямой и ориентацией его проекции:

$$\operatorname{tg}\alpha = \frac{1}{k}\operatorname{tg}A, \qquad (7.55)$$

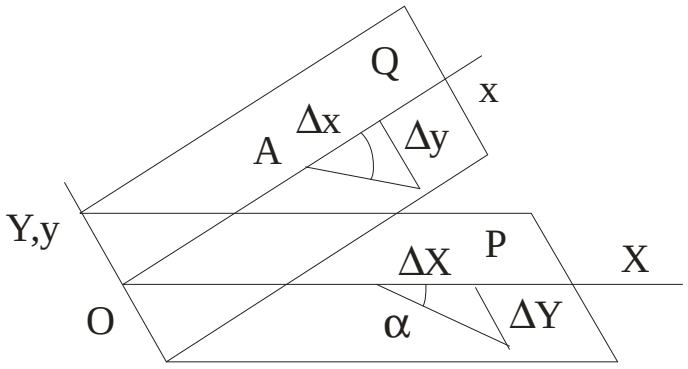


Рис. 7.9. Проекция угла

где A – угол между отрезком прямой и осью Ox на плоскости Q, а α – угол между проекцией отрезка и осью OX в плоскости P, т. е. α является проекцией угла A (рис. 7.9).

На основании последнего выражения можно получить другие соотношения между функциями углов A и α, например:

$$\left.\begin{aligned}\sin\alpha &= \frac{\operatorname{tg}A}{\sqrt{k^2+\operatorname{tg}^2 A}}\\ \cos\alpha &= \frac{k}{\sqrt{k^2+\operatorname{tg}^2 A}}\end{aligned}\right\} \qquad (7.56)$$

или

$$\left.\begin{aligned}\sin\alpha &= \frac{\sin A}{\mu}\\ \cos\alpha &= \frac{k\cos A}{\mu}\end{aligned}\right\}, \qquad (7.57)$$

где

$$\mu = \sqrt{\sin^2 A + k^2\cos^2 A}\,. \qquad (7.58)$$

Из данной формулы путем несложных преобразований можно получить, например, выражения:

$$\mu = \sqrt{k^2 + (1-k^2)\sin^2 A};$$

$$\mu = \sqrt{1-(1-k^2)\cos^2 A}$$

и ряд других производных формул. Однако формулу (7.58) будем считать основной по причине ее большей очевидности и, чтобы показать ее

структуру, далее не будем приводить ее к более удобному для вычислений виду.

Если известно значение угла α, то для вычисления значения коэффициента μ может использоваться основная формула

$$\mu = \frac{k}{\sqrt{\cos^2\alpha + k^2 \sin^2\alpha}} \qquad (7.59)$$

и выводимые из нее соотношения

$$\mu = \frac{k}{\sqrt{1-(1-k^2)\sin^2\alpha}};$$

$$\mu = \frac{k}{\sqrt{k^2+(1-k^2)\cos^2\alpha}}$$

и другие им подобные.

Кроме того, могут оказаться полезными формулы, выражающие углы как функции параметра μ:

$$\left.\begin{aligned} \sin A &= \sqrt{\frac{\mu^2-k^2}{1-k^2}} \\ \cos A &= \sqrt{\frac{1-\mu^2}{1-k^2}} \\ \mathrm{tg}A &= \sqrt{\frac{\mu^2-k^2}{1-\mu^2}} \end{aligned}\right\} \qquad (7.60)$$

и

$$\left.\begin{aligned} \sin\alpha &= \frac{1}{\mu}\sqrt{\frac{\mu^2-k^2}{1-k^2}} \\ \cos\alpha &= \frac{k}{\mu}\sqrt{\frac{1-\mu^2}{1-k^2}} \\ \mathrm{tg}\alpha &= \frac{1}{k}\sqrt{\frac{\mu^2-k^2}{1-\mu^2}} \end{aligned}\right\}. \qquad (7.61)$$

Зависимость между длиной произвольно ориентированного отрезка d и длиной его проекции D определяется выражением

$$D = \sqrt{\Delta X^2 + \Delta Y^2} = \mu\, d,$$

из которого следует, что μ представляет собой масштаб проекции отрезка прямой, имеющей направление A на исходной плоскости, поэтому его можно считать масштабом по направлению A:

$$\mu = \frac{D}{d}. \qquad (7.62)$$

Если сейчас, после выяснения смысла коэффициента μ, мы вернемся к выражению (7.58), то обнаружим, что масштаб μ равен отношению

$$\mu = \frac{\sin A}{\sin \alpha}. \qquad (7.63)$$

Данный факт становится более понятным, если формулу для вычисления μ вывести иначе.

Пусть d – расстояние между двумя произвольными точками на плоскости Q. Тогда справедливо равенство

$$\Delta y = d \sin A. \qquad (7.64)$$

На плоскости P расстояние между проекциями этих точек равно D и будет иметь место соотношение

$$\Delta Y = D \sin \alpha. \qquad (7.65)$$

Из последних двух выражений с учетом (7.54) следует равенство

$$\frac{D}{d} = \frac{\sin A}{\sin \alpha}, \qquad (7.66)$$

в котором отношение в левой части по определению представляет собой масштаб μ.

В качестве вспомогательной величины введем значение масштаба η по направлению, ортогональному к направлению A:

$$\eta(A) = \mu(A + 90). \qquad (7.67)$$

Тогда для вычисления значения η может использоваться формула

$$\eta = \sqrt{\cos^2 A + k^2 \sin^2 A}, \qquad (7.68)$$

а также ее варианты

$$\eta = \sqrt{k^2 + (1 - k^2)\cos^2 A};$$

$$\eta = \sqrt{1 - (1 - k^2)\sin^2 A}$$

и другие, получаемые из (7.58) заменой функций угла α на кофункции.

Из равенств (7.58) и (7.68) следует, что сумма квадратов масштабов по любым двум ортогональным на плоскости Q направлениям является константой:

$$\mu^2 + \eta^2 = 1 + k^2. \qquad (7.69)$$

Из данного равенства с учетом (7.58) получаем формулу для вычисления значения η по известному значению угла α:

$$\eta = \sqrt{\frac{\cos^2\alpha + k^4\sin^2\alpha}{\cos^2\alpha + k^2\sin^2\alpha}} = \sqrt{\frac{k^4 + (1-k^4)\cos^2\alpha}{k^2 + (1-k^2)\cos^2\alpha}} = \sqrt{\frac{1-(1-k^4)\sin^2\alpha}{1-(1-k^2)\sin^2\alpha}}\ ,$$

(7.70)

где возможна любая комбинация выражения в числителе с выражением в знаменателе.

Продифференцировав (7.58), получим выражение

$$\frac{d\mu}{dA} = \frac{(1-k^2)\sin A\cos A}{\sqrt{\sin^2 A + k^2\cos^2 A}}, \qquad (7.71)$$

из которого следует, что масштаб μ принимает экстремальные значения при $A = 0°$ и $A = 90°$.

Значение масштаба по оси абсцисс равно значению коэффициента k:

$$\mu_X = k. \qquad (7.72)$$

Данное выражение для масштаба по оси OX может быть получено непосредственно из формулы (7.52). Значение масштаба по оси ординат не зависит от угла ε между плоскостями и всегда равно 1:

$$\mu_Y = 1, \qquad (7.73)$$

что следует и из равенства (7.53). Отношение экстремальных значений масштаба является константой:

$$\frac{\mu_X}{\mu_Y} = k\,. \qquad (7.74)$$

Масштаб η будет иметь экстремальные значения при тех же значениях угла A, так как направление масштаба η ортогонально направлению масштаба μ, а направления экстремальных значений μ также ортогональны. Если значение масштаба η известно, то для вычисления значений функций углов A и α могут быть использованы формулы

$$\left.\begin{aligned} \sin A &= \sqrt{\frac{1-\eta^2}{1-k^2}} \\ \cos A &= \sqrt{\frac{\eta^2-k^2}{1-k}} \\ \mathrm{tg}A &= \sqrt{\frac{1-\eta^2}{\eta^2-k^2}} \end{aligned}\right\} \qquad (7.75)$$

и

$$\left.\begin{aligned}\sin\alpha &= \sqrt{\frac{1-\eta^2}{(1-k^2)(1+k^2-\eta^2)}}\\ \cos\alpha &= k\sqrt{\frac{\eta^2-k^2}{(1-k^2)(1+k^2-\eta^2)}}\\ \operatorname{tg}\alpha &= \frac{1}{k}\sqrt{\frac{1-\eta^2}{\eta^2-k^2}}\end{aligned}\right\}. \qquad (7.76)$$

Из сопоставления (7.63) и (7.69) можно получить формулы, выражающие связь между A и α через η:

$$\left.\begin{aligned}\frac{\sin A}{\sin\alpha} &= \sqrt{1+k^2-\eta^2}\\ \frac{\cos A}{\cos\alpha} &= \frac{1}{k}\sqrt{1+k^2-\eta^2}\end{aligned}\right\}. \qquad (7.77)$$

Введем на плоскости Q систему полярных координат с полюсом в точке O и полярной осью, совпадающей с осью Ox. Полярный угол u и полярный радиус r связаны с прямоугольными координатами x и y уравнениями:

$$\left.\begin{aligned}x &= r\cos u\\ y &= r\sin u\end{aligned}\right\} \qquad (7.78)$$

и

$$\left.\begin{aligned}r &= \sqrt{x^2+y^2}\\ \operatorname{tg}u &= \frac{y}{x}\end{aligned}\right\}. \qquad (7.79)$$

Если на картинной плоскости аналогичным образом введем систему полярных координат φ и ρ, то зависимость между полярными и прямоугольными координатами на плоскости проектирования будет выражаться аналогичными по виду формулами.

Радиус-вектору произвольной точки m, лежащей на плоскости Q и имеющей полярные координаты u и r, будет соответствовать радиус-вектор ее проекции M с полярными координатами φ и ρ. Формулы (7.52)–(7.79) устанавливают зависимость между геометрическими величинами на исходной плоскости Q и их проекциями на плоскости P. Поэтому соотношения между полярными координатами на плоскости Q и полярными координатами на плоскости проектирования легко могут быть получены путем замены в выражениях переменных A, d, α и D соответственно на u, r, φ и ρ и здесь не приводятся.

Значение коэффициента k в общем случае зависит как от угла ε, так и от угла ν. При $\varepsilon = 0$, как следует из (7.50), $k = 1$ и отображение является тривиальным, поэтому угол ε между плоскостями будем считать некоторой константой, не равной 0. Тогда, изменяя значение угла ν, можно получить различные проекции плоскости Q на картинную плоскость. Для наших целей достаточно рассмотреть отображение положительной полуплоскости $Q+$ на положительную полуплоскость $P+$. При этом характерными будут пять значений угла ν.

$\nu = -90°$. Это граничное значение, при котором направление проектирования параллельно оси Ox. Значение коэффициента $k = 0$, и все точки, лежащие на прямой OA, проектируются в точку O, а вся плоскость Q отображается на ось OY.

$\nu = 90°$. Это второе граничное значение. Направление проектирования параллельно оси OX, значение коэффициента $k = \infty$. Все точки прямой OA проектируются в бесконечность.

$\nu = -\varepsilon/2$. С изменением направления проектирования ν от -90° до 90° масштаб проекции отрезка, параллельного оси Ox, изменяется непрерывным образом от 0 до ∞. Следовательно, при некотором значении угла ν коэффициент k примет значение, равное 1. Из равенства

$$k = \frac{\cos\nu}{\cos(\varepsilon + \nu)} = 1 \qquad (7.80)$$

находим, что значение угла ν будет равно $-\varepsilon/2$. Направление проектирования при этом будет ортогонально биссектрисе угла ε, а плоскость проектирования будет представлять собой точную копию исходной плоскости.

Перечисленные случаи не представляют интереса по очевидным причинам. Наиболее важными частными случаями, представляющими практический интерес, являются значения угла $\nu = 0$ и $\nu = -\varepsilon$ (рис. 7.10).

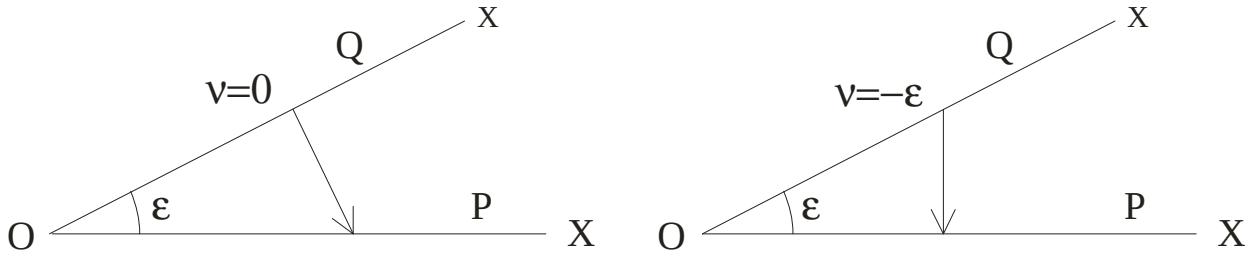


Рис. 7.10. Направления проектирования

В обоих случаях, как при $\nu = 0$, так и при $\nu = \varepsilon$, в результате проектирования исходной плоскости на картинную плоскость окружность трансформируется в эллипс с эксцентриситетом $e = \sin\varepsilon$, то есть форма эллипса сохраняется. Разница между эллипсами – в их размерах и ориентации относительно системы координат на картинной плоскости.

При $\nu = 0$ направление проектирования ортогонально исходной плоскости, картинная плоскость растянута в $k = 1/\cos\varepsilon$ раз по оси OX, малая полуось эллипса b равна радиусу окружности, и он вытянут вдоль оси абсцисс. При

$\nu = -\varepsilon$ направление проектирования ортогонально картинной плоскости, она сжата в k раз по оси абсцисс, и значение коэффициента k равно $\cos \varepsilon$. Эллипс развернут на 90°, а его большая полуось равна радиусу окружности. Поэтому первый эллипс выше был назван большим, а второй – малым.

Таким образом, свойства эллипса как геометрической фигуры определяются свойствами отображения, его растяжения или сжатия по одному направлению (оси OX) относительно исходной плоскости. Случай сжатия мы рассматривать не будем по причине его симметрии. Можно получить сонм формул для малого эллипса, если в формулах для большого эллипса заменить $1/\cos\varepsilon$ на $\cos\varepsilon$.

Теперь рассмотрим формулы, определяющие значения масштабов μ и η. Как видно из (7.58), значение μ зависит только от направления A, поэтому μ определяет масштаб не только вдоль радиус-вектора, но и вдоль любой прямой, параллельной ему. На рис. 7.9 была рассмотрена проекция произвольного отрезка и угла между ним и осью Ox. Теперь обратимся к важному частному случаю, когда такой отрезок является радиусом окружности (рис. 7.11).

Очевидно, что угол на плоскости Q, который мы раньше обозначали как A, будет широтой u точки на сфере, а соответствующий ему угол β на плоскости P – геоцентрической широтой φ точки на эллипсе, являющейся проекцией выбранной точки на окружности. Соотношение между ними определяется как частный случай формулы (7.55):

$$\operatorname{tg}\varphi = \frac{1}{k}\operatorname{tg}u. \qquad (7.81)$$

Так как $\frac{1}{k} = \cos\varepsilon$ и $\cos\varepsilon = \sqrt{1-e^2}$, то в традиционных обозначениях формула (7.81) примет привычный вид

$$\operatorname{tg}\varphi = \sqrt{1-e^2}\ \operatorname{tg}u.$$

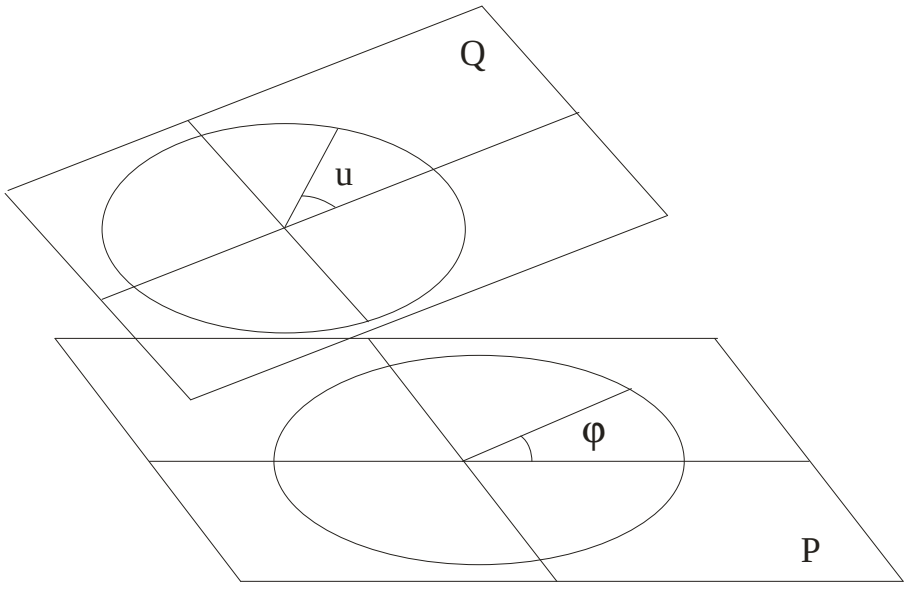


Рис. 7.11. К определению масштабов проекции

Полученное выражение есть известное соотношение между приведенной и геоцентрической широтой точки. Таким образом, *приведенная широта*, интерпретацию которой выше мы отложили, представляет собой широту точки на окружности, принадлежащей исходной плоскости *Q*. Иначе – геоцентрическая широта φ является проекцией приведенной широты u.

Формулы для вычисления масштабов μ и η также остаются справедливыми:

$$\mu = \sqrt{k^2 \cos^2 u + \sin^2 u}; \qquad (7.82)$$

$$\eta = \sqrt{\cos^2 u + k^2 \sin^2 u}. \qquad (7.83)$$

Значения масштабов могут быть выражены через значение геоцентрической широты, для чего достаточно воспользоваться формулами (7.59) и (7.69):

$$\mu = \frac{k}{\sqrt{\cos^2 \varphi + k^2 \sin^2 \varphi}}; \qquad (7.84)$$

$$\eta = \sqrt{\frac{\cos^2 \varphi + k^4 \sin^2 \varphi}{\cos^2 \varphi + k^2 \sin^2 \varphi}}. \qquad (7.85)$$

В сфероидической геодезии чаще используется геодезическая широта. Значения масштабов могут быть представлены как ее функции

$$\mu = \sqrt{\frac{k^4 \cos^2 B + \sin^2 B}{k^2 \cos^2 B + \sin^2 B}}; \qquad (7.86)$$

$$\eta = \frac{1}{\sqrt{1 - e^2 \sin^2 B}}. \qquad (7.87)$$

Выше говорилось, что μ определяет масштаб вдоль радиус-вектора, а η задает значение масштаба по касательной к окружности. Касательная к окружности отображается в касательную к эллипсу, поэтому η будет также определять значение масштаба вдоль дуги эллипса в точке с приведенной широтой u. По этой причине значение μ может быть названо *радиальным масштабом,* а η *– тангенциальным масштабом*.

Знаменатель формулы (7.87) представляет собой значение первой геодезической величины в традиционных обозначениях. Поэтому с учетом (7.41) можно записать

$$\eta = \frac{1}{W} \qquad (7.88.1)$$

и

$$\eta = \frac{k}{V}. \qquad (7.88.2)$$

Таким образом, мы установили, что первая геодезическая величина является величиной, обратной значению тангенциального масштаба η. Масштаб

служит фундаментальной характеристикой любого отображения, допускающего геометрическую интерпретацию. Эллипс является отображением окружности, поэтому нет ничего удивительного в том, что масштаб отображения (в завуалированном виде) встречается во многих формулах сфероидической геодезии.

Возможно, что дать геометрическую трактовку величин W и V при традиционном изложении курса сфероидической геодезии было бы трудно, поскольку не используются понятия линейного отображения и его масштаба. С проективной точки зрения таинственные обозначения W и V приобретают ясный геометрический смысл.

7.4. Вычисление длины дуги эллипса

Задача вычисления длины дуги эллипса возникает при решении некоторых геодезических задач. К ней, например, сводится задача вычисления абсцисс при определении координат в проекции Гаусса – Крюгера.

Если дугу окружности обозначить как s, а дугу эллипса – как S, то между бесконечно малым элементом dS дуги эллипса и бесконечно малым элементом ds дуги окружности будет иметь место соотношение

$$dS = \eta\, ds. \qquad (7.89)$$

Так как

$$ds = b\, du, \qquad (7.90)$$

то

$$dS = b\eta\, du. \qquad (7.91)$$

Чтобы выразить данное соотношение через геодезическую широту, используем зависимость между приведенной и геодезической широтами:

$$tgB = k\, tgu,$$

где $k = 1/\cos\varepsilon$.

Продифференцировав его, получим

$$\frac{dB}{\cos^2 B} = k\frac{du}{\cos^2 u}$$

или

$$du = \frac{\eta^2}{k}dB.$$

Подставив это значение в (7.91), приходим к выражению

$$dS = \frac{b}{k}\eta^3\, dB.$$

В геодезической литературе полученное соотношение обычно представляется как

$$dS = \frac{a(1-e^2)dB}{(1-e^2\sin^2 B)^{3/2}} = \frac{a(1-e^2)dB}{W^3}. \qquad (7.92)$$

Длина дуги меридиана от экватора до параллели с широтой B является не выражающимся с помощью конечной комбинации элементарных функций эллиптическим интегралом

$$S = \int_0^B \frac{a(1-e^2)dB}{\sqrt{(1-e^2\sin^2 B)^3}}. \qquad (7.93)$$

Традиционным приемом ее вычисления является разложение интеграла (7.93) в ряд и почленное интегрирование, обеспечивающее любую необходимую точность [4], [7], [8]. Для достижения наибольшей практически необходимой точности 0.001 м в разложении удерживаются члены, содержащие e^6 и функции кратных углов до $6B$ включительно. Типичным примером такого разложения является выражение, приведенное в [8, с. 39]:

$$x_0 = n_1 B_0 - n_2 \sin 2B_0 + n_3 \sin 4B_0 - n_4 \sin 6B_0 + \ldots, \qquad (7.94)$$

где

$$\left.\begin{aligned} n_1 &= c(1-\frac{3}{4}e'^2+\frac{45}{64}e'^4-\frac{175}{256}e'^6+\frac{11025}{16384}e'^8-\cdots)\\ n_2 &= \frac{3}{8}c\cdot e'^2\,(1-\frac{5}{4}e'^2+\frac{175}{128}e'^4-\frac{105}{64}e'^6+\cdots)\\ n_3 &= \frac{15}{256}c\cdot e'^4\,(1-\frac{7}{4}e'^2+\frac{147}{64}e'^4-\cdots)\\ n_4 &= \frac{35}{3072}c\cdot e'^6\,(1-\frac{9}{4}e'^2+\cdots) \end{aligned}\right\}. \qquad (7.95)$$

В [7] для вычисления длины дуги меридиана предлагается общее выражение

$$S = a_0 B - \frac{a_2}{2}\sin 2B + \frac{a_4}{4}\sin 4B - \frac{a_6}{6}\sin 6B + \frac{a_8}{8}\sin 8B - \ldots, \qquad (7.96)$$

где a – коэффициенты, зависящие от параметров эллипса, последний из которых составляет всего лишь 0,03 мм и поэтому может быть отброшен. Вычисление дуги меридиана с ошибкой менее 0,1 мм для эллипсоида Красовского может осуществляться по формуле, приведенной в [7, с. 29]

$$S = 6\,367\,558.4969\,B -$$

$$-\sin B\cos B[32\,005.7801+(133.9213+0.7032\sin^2 B)\sin^2 B]. \qquad (7.97)$$

При больших расстояниях, когда $S/R>1$, подобные ряды начинают расходиться и тогда используют разложение по степеням e.

Применение ЭВМ для решения геодезических задач не снимает проблемы эффективности вычислений, поскольку эта проблема непреходяща и будет существовать всегда. Но разложение в ряды,

вероятно, слишком универсальное средство, чтобы быть столь же эффективным. Поэтому можно попытаться найти решение задачи вычисления длины дуги меридиана (или эллиптического интеграла), обеспечивающее необходимую практическую точность, но более эффективное, чем разложение в ряд. Решение поставленной задачи приводится ниже. В качестве исходного используем равенство (7.91) и тогда

$$S = b\int_0^u \eta\, du\,. \qquad (7.98)$$

Выше эллипс рассматривался как сечение прямого кругового цилиндра плоскостью (см. рис. 7.3), и была дана интерпретация первого эксцентриситета эллипса как $e = \sin\varepsilon$. Если развернуть поверхность цилиндра на плоскость, то дуга эллипса трансформируется в косинусоиду, описываемую уравнением

$$z = b\,\mathrm{tg}\varepsilon\cos u\,, \qquad (7.99)$$

которое может быть получено непосредственно из рис. 7.3, и в котором u – приведенная широта произвольной точки M. Тангенс угла наклона ν между касательной к эллипсу в произвольной точке и плоскостью окружности равен

$$\mathrm{tg}\,\nu = \frac{dz}{ds} = \frac{dz}{du}\frac{du}{ds} = -\mathrm{tg}\varepsilon\sin u\,. \qquad (7.100)$$

Зависимость между бесконечно малым элементом dS дуги эллипса и бесконечно малым элементом ds дуги окружности (рис. 7.12) определяется равенствами:

$$dS = \frac{ds}{\cos\nu} = \sqrt{1+\mathrm{tg}^2\nu}\; ds = \sqrt{1+\mathrm{tg}^2\varepsilon\sin^2 u}\; ds\,. \qquad (7.101)$$

Если предположить, что нам неизвестно уравнение кривой, длина которой представляется дифференциальным уравнением (7.101), то мы можем получить его на основании следующих соображений. Известно, что длина элементарного отрезка плоской кривой может быть выражена соотношением

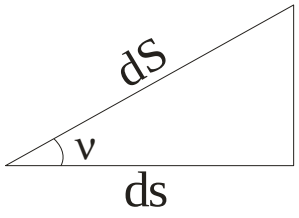


Рис. 7.12. Элементы дуг окружности и эллипса

$$dS = b\sqrt{1+\left(\frac{df}{du}\right)^2}\; du\,.$$

Тогда с учетом соотношения $ds = bdu$ мы можем написать равенство

$$\sqrt{1+\left(\frac{df}{du}\right)^2} = \sqrt{1+\mathrm{tg}^2\varepsilon\sin^2 u}\,,$$

из которого находим

$$\frac{df}{du} = \pm b \operatorname{tg}\varepsilon \sin u .$$

Проинтегрировав данное уравнение, получаем выражение неизвестной функции

$$f(u) = \pm b \operatorname{tg}\varepsilon \cos u + C ,$$

представляющее собой уравнение косинусоиды. Следовательно, определение длины дуги эллипса сводится к вычислению длины косинусоиды.

Для нахождения приближения эллиптического интеграла представим элемент *dS* как сумму

$$dS = ds + d\sigma . \qquad (7.102)$$

При малом значении эксцентриситета $\sin\varepsilon$ значение $d\sigma$ намного меньше *dS*. Идея приближения *dS* проста и заключается в том, что если мы найдем приближение $d\sigma$ с некоторой относительной ошибкой *m*, то относительная ошибка элемента *dS*, при равенстве абсолютных ошибок, будет намного меньше *m*.

С этой целью дополним вспомогательными построениями рис. 7.12 и получим рис. 7.13, на котором *ds* = *OA* = *OC* и $d\sigma$ = *CB* = *FB*. Из рис. 7.13 следует

$$d\sigma = \operatorname{tg}\nu \operatorname{tg}\frac{\nu}{2} ds . \qquad (7.103)$$

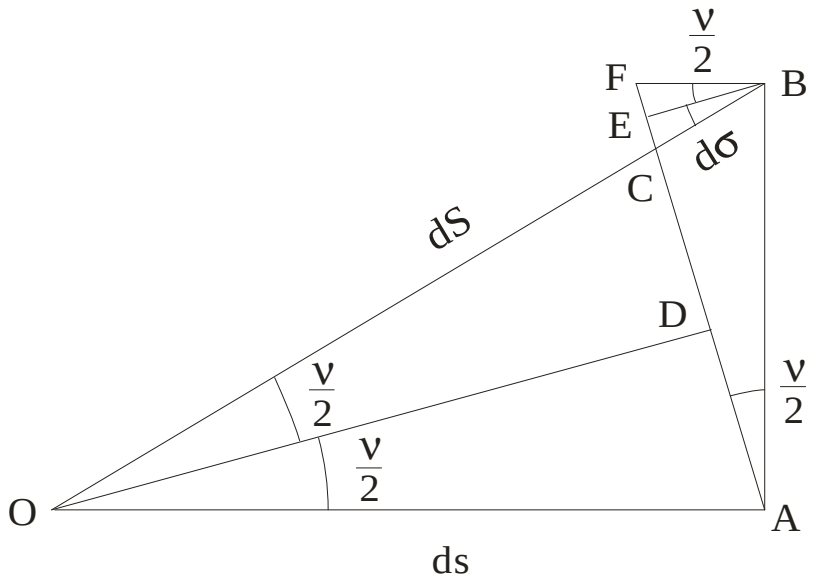


Рис. 7.13. Приращение dσ

Данное выражение можно получить и аналитическим путем:

$$d\sigma = dS - ds = \frac{ds}{\cos\nu} - ds = \frac{1-\cos\nu}{\cos\nu} ds = \frac{\sin\nu}{\cos\nu}\frac{1-\cos\nu}{\sin\nu} ds = \operatorname{tg}\nu \operatorname{tg}\frac{\nu}{2} ds .$$

Интеграл

$$S = \int_0^u (ds + d\sigma) = b\int_0^u du + b\int_0^u \operatorname{tg}\nu \operatorname{tg}\frac{\nu}{2} du , \qquad (7.104)$$

где $\operatorname{tg}\nu = -\operatorname{tg}\varepsilon \sin u$, по-прежнему сводится к эллиптическому.

Известное из тригонометрии соотношение

$$tg\nu = \frac{2\mathrm{tg}\frac{\nu}{2}}{1-\mathrm{tg}^2\frac{\nu}{2}} \qquad (7.105)$$

представим в виде

$$\mathrm{tg}\frac{\nu}{2} = \frac{1}{2}\mathrm{tg}\,\nu(1-\mathrm{tg}^2\frac{\nu}{2}) \qquad (7.106)$$

и, подставив последнее выражение в (7.103), получим

$$d\sigma = \frac{1}{2}\mathrm{tg}^2\nu(1-\mathrm{tg}^2\frac{\nu}{2})\,ds. \qquad (7.107)$$

Введем еще одну подстановку

$$\mathrm{tg}\frac{\nu}{2} = \frac{1}{q}\mathrm{tg}\,\nu, \qquad (7.108)$$

где q – некоторый коэффициент.

При малых углах значение тангенса половинного угла часто принимают равным половине тангенса полного угла. В формуле (7.107) такое приближение может оказаться грубым, поскольку абсолютное значение угла ν, как следует из (7.100), изменяется в пределах от 0 до ε, что для любого земного эллипсоида (для эллипсоида Красовского $\varepsilon = 4°41'34{,}093\,903\,7''$) уже может оказаться заметной величиной. Погрешность такого приближения будет возрастать по мере увеличения угла наклона ν и достигнет наибольшего значения при $u = 90°$. Более точным будет значение q, вычисленное для максимального по модулю значения угла наклона ν:

$$q = \frac{\mathrm{tg}\varepsilon}{\mathrm{tg}\frac{\varepsilon}{2}}. \qquad (7.109)$$

К этому значению мы вернемся далее, а пока продолжим наш вывод в общем виде. На основании (7.106) и (7.108) можно записать

$$\mathrm{tg}^2\frac{\nu}{2} = \frac{1}{2q}\mathrm{tg}^2\nu(1-\mathrm{tg}^2\frac{\nu}{2}). \qquad (7.110)$$

Подставив последнее выражение в (7.107), получим

$$d\sigma = \frac{1}{2}\mathrm{tg}^2\nu(1-\frac{1}{2q}\mathrm{tg}^2\nu(1-\mathrm{tg}^2\frac{\nu}{2}))ds. \qquad (7.111)$$

Применяя к данному выражению подстановку (7.110) неограниченное число раз, приходим к соотношению

$$d\sigma = \frac{1}{2}\mathrm{tg}^2\nu(1-\frac{1}{2q}\mathrm{tg}^2\nu(1-\frac{1}{2q}\mathrm{tg}^2\nu(1-\frac{1}{2q}\mathrm{tg}^2\nu(1-...))))\,ds \qquad (7.112)$$

или

$$d\sigma = (\frac{1}{2}\operatorname{tg}^2\nu - \frac{1}{4q}\operatorname{tg}^4\nu + \frac{1}{8q^2}\operatorname{tg}^6\nu - \frac{1}{16q^3}\operatorname{tg}^8\nu + ...)\, ds. \qquad (7.113)$$

Последовательность в скобках можно считать бесконечной геометрической прогрессией с первым членом

$$a = \frac{1}{2}\operatorname{tg}^2\nu \qquad (7.114)$$

и знаменателем прогрессии

$$z = -\frac{1}{2q}\operatorname{tg}^2\nu, \qquad (7.115)$$

сумма членов которой равна

$$\Sigma = \frac{a}{1-z}. \qquad (7.116)$$

При малых углах наклона ν знаменатель прогрессии меньше 1, и при неограниченном возрастании числа ее членов последний член стремится к 0. Поэтому на основании формулы для вычисления суммы членов бесконечной геометрической прогрессии:

$$d\sigma = \frac{\frac{1}{2}\operatorname{tg}^2\nu}{1+\frac{1}{2q}\operatorname{tg}^2\nu} ds = b\frac{\frac{1}{2}\operatorname{tg}^2\varepsilon\sin^2 u}{1+\frac{1}{2q}\operatorname{tg}^2\varepsilon\sin^2 u} du. \qquad (7.117)$$

От $\sin u$ в числителе можно избавиться, если дополнить последовательность в (7.113) слева двумя членами с противоположными знаками

$$d\sigma = (q - q + \frac{1}{2}\operatorname{tg}^2\nu - \frac{1}{4q}\operatorname{tg}^4\nu + \frac{1}{8q^2}\operatorname{tg}^6\nu - \frac{1}{16q^3}\operatorname{tg}^8\nu + ...)\, ds. \qquad (7.118)$$

После перегруппировки данного выражения получим

$$d\sigma = q\, ds - (q - \frac{1}{2}\operatorname{tg}^2\nu + \frac{1}{4q}\operatorname{tg}^4\nu - \frac{1}{8q^2}\operatorname{tg}^6\nu + \frac{1}{16q^3}\operatorname{tg}^8\nu - ...)\, ds. \qquad (7.119)$$

Выражение в скобках является геометрической прогрессией с тем же знаменателем прогрессии, но другим первым членом:

$$a = q. \qquad (7.120)$$

Сумма ее членов равна

$$\Sigma = \frac{q}{1+\frac{1}{2q}\operatorname{tg}^2\nu}. \qquad (7.121)$$

Поэтому выражение (7.119) преобразуется к виду

$$d\sigma = q\,ds - \frac{q}{1+\frac{1}{2q}\operatorname{tg}^2\nu}\,ds. \qquad (7.122)$$

Можно показать, что полученное выражение может быть преобразовано к (7.103), следовательно, оно является точным. Но мы привели его к виду, в котором зависящий от u коэффициент q можно считать константой, поскольку угол ν изменяется в небольших пределах (от 0 до 4,7°). В итоге мы получаем следующую формулу для вычисления длины дуги эллипса:

$$S = \int_0^u dS = \int_0^u ds + \int_0^u d\sigma = b\int_0^u du + b\,q\int_0^u du - b\,q\int_0^u \frac{du}{1+\frac{1}{2q}\operatorname{tg}^2\varepsilon\sin^2 u}. \qquad (7.123)$$

Если q считать константой, то для интегрирования (7.123) можно воспользоваться интегралом из [3, с. 166]:

$$\int \frac{dx}{a+b\sin^2 x} = \frac{sign(a)}{\sqrt{a(a+b)}}\operatorname{arctg}\left(\sqrt{\frac{a+b}{a}}\operatorname{tg}x\right)\ \left[\frac{b}{a} > -1\right]. \qquad (7.124)$$

После интегрирования (7.123) находим

$$S = b\left((1+q)\,u - \frac{q}{\sqrt{1+\frac{1}{2q}\operatorname{tg}^2\varepsilon}}\operatorname{arctg}\left(\sqrt{1+\frac{1}{2q}tg^2\varepsilon}\ \operatorname{tg}u\right)\right). \qquad (7.125)$$

До интегрирования мы рассуждали строго, но при интегрировании допустили прегрешение, считая q константой. Отсюда следует, что формула (7.125) является приближенной, и ее точность зависит от характера изменения коэффициента q, являющегося функцией угла ν, что видно из (7.108). Значение ν определяется равенством (7.100), из которого следует, что при изменении широты точки от 0 до 90° угол ν изменяется в пределах от 0 до ε. Для эллипсоида Красовского точные значения коэффициента q в некоторых точках приводятся в табл. 7.1, а на рис. 7.14 представлен график изменения q и для сравнения – $\operatorname{tg}\nu$. На графике видно, что изменение $\operatorname{tg}\nu$ близко к линейному закону, но коэффициент q, хотя и изменяется в небольшом диапазоне, заметно отличается от линейной функции.

Таблица 7.1. Значения коэффициента q

ν	$\mathrm{tg}\nu$	$\mathrm{tg}\frac{\nu}{2}$	$q=\frac{\mathrm{tg}\,\nu}{\mathrm{tg}\,\frac{\nu}{2}}$
0			2,000 000 000 000 000
$\frac{1}{4}\varepsilon$	0,020 479 081 843 904	0,010 238 467 549 606	2,000 209 674 414 916
$\frac{1}{2}\varepsilon$	0,040 975 348 453 648	0,020 479 081 843 904	2,000 839 137 514 562
$\frac{3}{4}\varepsilon$	0,061 506 042 325 039	0,030 723 991 428 436	2,001 889 711 117 194
ε	0,082 088 521 826 432	0,040 975 348 453 648	2,003 363 605 785 883

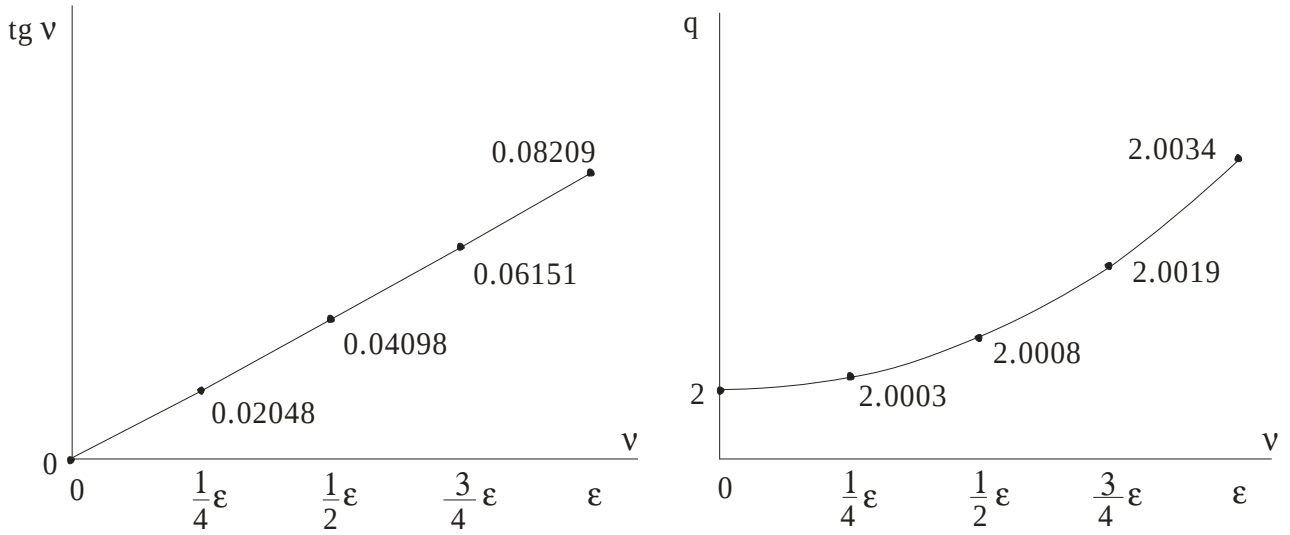


Рис. 7.14. График изменения q

Пока что мы предполагаем, возможно, без достаточных оснований, что использование равенства (7.109) обеспечит удовлетворительную точность вычисления длины дуги меридиана. Тогда мы можем привести формулу (7.125) к более простому виду. Если (7.109) позволяет достичь нужной точности, то подкоренное выражение в (7.125) допустимо считать константой для конкретного эллипса, которую достаточно вычислить один раз, поэтому можно ввести обозначение

$$p=\sqrt{1+\frac{1}{2q}\mathrm{tg}^2\varepsilon}\,. \qquad (7.126)$$

Из (7.126) и допустимости (7.109) следует

$$p=\sqrt{\frac{q}{2}}\,, \qquad (7.127)$$

что можно записать также как

$$2p^2 = q. \qquad (7.128)$$

Тогда (7.125) можно представить в виде

$$S = b((1+q)u - 2p\,\mathrm{arctg}(p\,\mathrm{tg}u)). \qquad (7.129)$$

Но из (7.109) также следует короткое

$$q = k + 1, \qquad (7.130)$$

в котором $k = 1/\cos\varepsilon$. Подставив значение q в (7.129), приходим к выражению

$$S = b((2+k)u - 2p\,\mathrm{arctg}(p\,\mathrm{tg}u)). \qquad (7.131)$$

Полученный результат можно еще упростить, если ввести обозначения константных выражений

$$b(2+k) = p_1; \qquad (7.132)$$

$$2bp = p_2. \qquad (7.133)$$

В результате получаем лаконичное выражение

$$S = p_1 u - p_2\,\mathrm{arctg}(p\,\mathrm{tg}u). \qquad (7.134)$$

Таким образом, для получения длины дуги меридиана от экватора до параллели с приведенной широтой *u* необходимо вычислить коэффициенты:

1) 1) $p = \sqrt{\frac{k+1}{2}}$;

2) 2) $p_1 = b(2+k)$;

3) 3) $p_2 = 2bp$

– и применить формулу (7.134).

Коэффициенты p, p_1 и p_2 , зависящие только от размеров и формы эллипса, могут считаться такими же параметрами эллипса, как его полуоси, эксцентриситет или сжатие, и могут быть определены для конкретного земного эллипсоида один раз.

Для оценки точности полученной формулы (7.134) были вычислены длины дуг меридиана с интервалом 10° и выполнено их сравнение с «точными» значениями. Результаты вычисления приведены в табл. 7.2. Для вычисления «точного» значения было использовано разложение эллиптического интеграла в ряд (7.97). Приближенные значения длины дуги меридиана были получены при значении коэффициента q, вычисленного по формуле (7.109), и равного $q = 2{,}003363605786$. В последней графе табл. 7.2 указаны отклонения приближенных значений от точных, выраженные в миллиметрах.

Из формулы (7.134) можно вывести ее различные варианты, например, содержащие обратные тригонометрические функции arcsin или arccos:

$$S = p_1 u - p_2 \arccos\frac{\cos u}{\sqrt{\cos^2 u + p^2 \sin^2 u}}; \qquad (7.135)$$

$$S = p_1 u - p_2 \arcsin \frac{p \sin u}{\sqrt{\cos^2 u + p^2 \sin^2 u}}. \qquad (7.136)$$

Таблица 7.2. Разности между приближенным и точным значениями длины дуги меридиана

Широта B°	**Значение длины дуги меридиана**		Погрешность (мм)
	Точное (м)	Приближенное (м)	
0	0, 000 0	0, 000	0
10	1 105 874,609 4	1 105 874,609 4	0,0
20	2 212 405,724 3	2 212 405,724 3	0,0
30	3 320 172,406 8	3 320 172,407 1	+0,3
40	4 429 607,367 8	4 429 607,368 9	+1,1
50	5 540 944,467 6	5 540 944,470 0	+2,4
60	6 654 189,092 2	6 654 189,096 2	+4,0
70	7 769 115,633 7	7 769 115,638 8	+5,1
80	8 885 293,251 6	8 885 293,257 3	+5,7
90	10 002 137,497 6	10 002 137,503 5	+5,9

Полученные формулы позволяют вычислять длину дуги земного эллипсоида с относительной ошибкой менее 7×10^{-10}. При малых расстояниях выражения (7.134)–(7.136) являются практически точными. Например, ошибка вычисления дуги меридиана длиной 25 км будет менее 0.02 мм. С уменьшением эксцентриситета эллипса ошибка вычисления длины его дуги убывает. Для окружности $\varepsilon = 0$, поэтому полученные формулы превращаются в точные.

Основной целью получения предлагаемых формул было решение геометрических задач в среде геоинформационных систем. Полученные формулы могут применяться при обработке геодезических сетей. В [4, с. 88] отмечается: «… принимая ошибку во взаимном положении смежных пунктов триангуляции по осям координат в 6 см, мы должны потребовать, чтобы ошибки вычисления разности широт, долгот исходного и определяемого пунктов находились в пределах 0,6–1,0 см». Как следует из таблицы, максимальное значение погрешности вычисления длины дуги меридиана по полученным формулам не превышает указанного допуска.

Если обратиться к формулам (7.134–7.136), то очевидно, что точность вычисления длины дуги эллипса зависит от точности определения коэффициента q. Отклонения от точных значений можно сделать меньше 0.001 м, если ввести поправочный член δ. Для нахождения δ можно указать два способа. Первый способ – практичный, но не слишком привлекательный с эстетической точки зрения, заключается в вычислении таблицы поправок через 5 или 10 градусов широты и в вычислении поправочного члена линейным интерполированием между табличными значениями. Второе решение – представить q как некоторую простую функцию от широты.

На рис. 7.15 показан график отклонений приближенных значений от «точных». Распределение погрешностей близко к функции, имеющей вид

$y = \sin^3 x$, график которой показан на этом же рисунке. Поэтому точность формулы (7.134) можно повысить, если ее дополнить членом

$$\delta = -\Delta \sin^3 u,$$

где $\Delta = 5{,}9$ мм. Тогда максимальная погрешность приближения составит 0,5 мм при u = 40°, но при этом (7.134) теряет свою лаконичность. По мнению автора, формула (7.134) пригодна в любых системах геомоделирования, возможно, даже при решении задач высшей геодезии на большие расстояния.

Вычисление длин дуг меридианов более 90° по полученным формулам в пояснениях не нуждается в силу своей очевидности.

Интеграл (7.123) служит хорошим приближением эллиптического интеграла и может быть назван *квазиэллиптическим*. В качестве особенности его вывода можно отметить тот факт, что он был получен без использования радиуса кривизны эллипса.

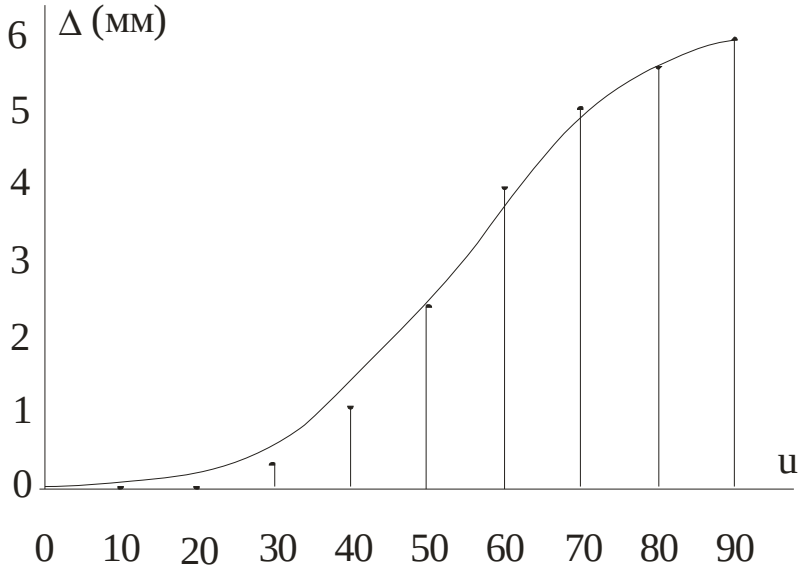


Рис. 7.15. График погрешностей

7.5. Сфера

Сфера – центральная поверхность второго порядка, уравнение которой в прямоугольной системе координат может быть представлено как

$$(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2 = R^2, \qquad (7.137)$$

где x_0, y_0, z_0 – координаты точки, называемой *центром сферы*, а R – *радиус сферы*. Если начало системы координат совпадает с центром сферы, то уравнение сферы принимает более простой вид:

$$x^2 + y^2 + z^2 = R^2. \qquad (7.138)$$

Пересечение любой плоскости со сферой образует окружность. Если секущая плоскость при этом проходит через центр сферы, то сечение называют *большим кругом* (рис. 7.16). Через две диаметрально противоположные точки на сфере можно провести бесконечное число больших кругов. Через любые две точки сферы, не являющиеся диаметрально противоположными, можно

провести единственный большой круг (дуга *AmB* на рис. 7.16). Дуги больших кругов на сфере играют такую же роль, как и прямые на плоскости. Их отличие от прямых на плоскости заключается в том, что дуга большого круга является кратчайшим расстоянием между двумя точками сферы только тогда, когда она меньше дополнительной дуги; дополнительная дуга в таком случае будет наибольшим расстоянием на сфере между этими точками.

Другое отличие дуг больших кругов от прямых на плоскости состоит в том, что на сфере не существует параллельных больших кругов и два больших круга всегда пересекаются (в двух точках).

Дуга *AB* большого круга измеряется соответствующим центральным углом *AOB* (рис. 7.16). *Угол на сфере*, образованный двумя большими кругами, измеряется углом между касательными к этим дугам в точке их пересечения или двугранным углом между соответствующими плоскостями (рис. 7.17). Если два больших круга пересекаются в некоторой точке на сфере, то они пересекаются и в диаметрально противоположной точке сферы.

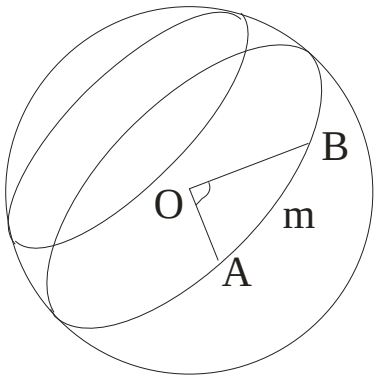


Рис. 7.16. Дуга большого круга

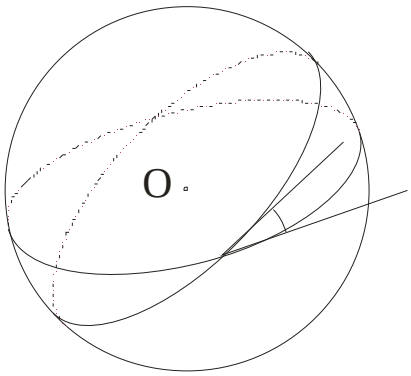


Рис. 7.17. Угол на сфере

При пересечении двух больших кругов на сфере образуется четыре *сферических двуугольника*, а при пересечении трех больших кругов, не пересекающихся в одной точке, образуется восемь *сферических треугольников* (рис. 7.18). Если известны элементы одного такого треугольника, то по ним могут быть определены элементы всех остальных треугольников. Поэтому обычно рассматривают соотношения между элементами только *эйлерова треугольника* – треугольника, все стороны которого меньше половины дуги большого круга (рис. 7.19).

В сферическом треугольнике его стороны измеряют соответствующими центральными углами. Свойства сферических треугольников имеют отличия от треугольников на плоскости. Так, два плоских треугольника равны, если:

1) любые две стороны одного треугольника и какой-либо из углов равны соответствующим сторонам и углу другого треугольника;

2) любые два угла и любая из сторон одного треугольника равны соответствующим двум углам и стороне другого треугольника;

3) три стороны одного треугольника равны трем сторонам другого.

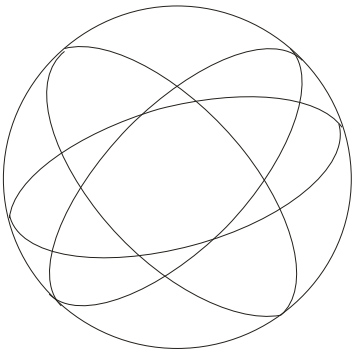


Рис. 7.18. Сферические треугольники

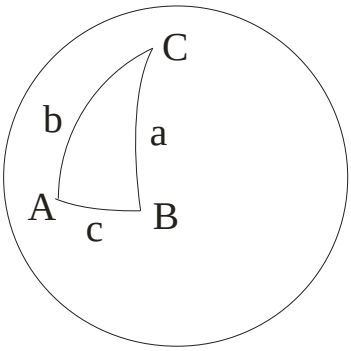


Рис. 7.19. Эйлеров треугольник

Для сферических треугольников справедливы указанные условия равенства и добавляется четвертое: два треугольника равны, если равны их углы. На сфере отсутствуют подобные треугольники: два сферических треугольника либо равны, либо не равны. Два сферических треугольника считаются равными, если они могут быть совмещены в результате их перемещения по сфере. Следовательно, сферические треугольники *равны*, если равны их элементы и треугольники имеют одинаковую ориентацию. Если сферические треугольники имеют одинаковые элементы, но противоположную ориентацию, то их называют *симметричными*.

В любом сферическом треугольнике каждая сторона меньше суммы и больше разности двух других сторон. Сумма сторон сферического треугольника всегда меньше 2π, а сумма углов больше π и меньше 3π. Предельным является сферический треугольник, все вершины которого лежат в плоскости одного большого круга, сумма его углов равна 3π, а сумма сторон 2π. Разность $\varepsilon = \Sigma - \pi$, где Σ – сумма углов сферического треугольника, называется *сферическим избытком*. Величина сферического избытка может быть вычислена по формуле

$$\varepsilon = \frac{S}{R^2}, \qquad (7.139)$$

где S – площадь сферического треугольника, а R – радиус сферы.

Положение любой точки на сфере может быть определено парой двух чисел, являющихся ее координатами. Система координат на сфере обычно вводится следующим образом. На сфере выбираются два взаимно перпендикулярных больших круга, один из которых называют *экватором*, а второй – *начальным*, или *нулевым меридианом*. Точки пересечения диаметра, перпендикулярного плоскости экватора, со сферой называют *полюсами* (точки P_1 и P_2 на рис. 7.20). Большие круги, проходящие через полюса, называют *меридианами*, а малые круги, параллельные плоскости экватора, – *параллелями*. В качестве одной координаты выбирается угол L между начальным меридианом и меридианом точки, называемый *долготой*. Другой координатой служит угол u между радиус-вектором точки и плоскостью экватора, называемый *широтой* точки (угол NOM на рис. 7.20). Иногда в качестве второй координаты используется *полярное расстояние* θ – дополнение широты до 90°. Меридианы

представляют собой координатные линии $L = const$, а параллели – $u = const$. Радиус r параллели с широтой u выражается формулой $r = R\cos u$.

Положение кривой на сфере задается ее уравнением $f(u, L)$ или двумя уравнениями в параметрической форме: $u = f(t)$ и $L = g(t)$, где t – некоторый параметр.

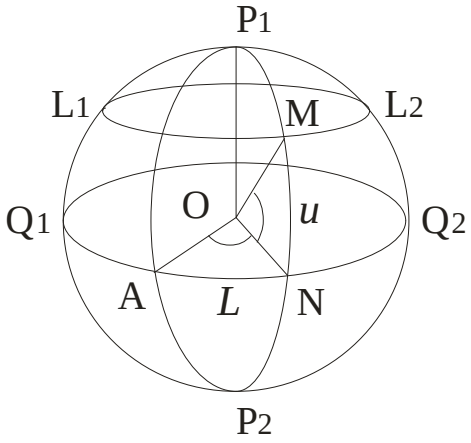


Рис. 7.20. Координаты на сфере

7.6. Формулы сферической тригонометрии

Зависимости между сторонами и углами сферических треугольников изучаются в сферической тригонометрии, возникшей задолго до появления «обычной» тригонометрии на плоскости. Основными соотношениями между элементами косоугольных сферических треугольников являются:

– формула синусов

$$\frac{\sin a}{\sin A} = \frac{\sin b}{\sin B} = \frac{\sin c}{\sin C}; \qquad (7.140)$$

– формула косинуса стороны

$$\cos a = \cos b \cos c + \sin b \sin c \cos A; \qquad (7.141)$$

– формула косинуса угла

$$\cos A = \cos B \cos C + \sin B \sin C \cos a; \qquad (7.142)$$

– формулы пяти элементов

$$\sin a \cos B = \cos b \sin c - \sin b \cos c \cos A; \qquad (7.143)$$

$$\sin A \cos b = \cos B \sin C - \sin B \cos C \cos a; \qquad (7.144)$$

– формула котангенсов

$$\mathrm{ctg}a \sin b = \cos b \cos C + \sin C\, ctgA. \qquad (7.145)$$

Для решения косоугольных треугольников может применяться *формула тангенса половинного угла*, называемая также *формулой полупериметра*:

$$\mathrm{tg}\frac{A}{2} = \frac{M}{\sin(p-a)}, \qquad (7.146)$$

где

$$p = \frac{1}{2}(a+b+c);$$

$$M = \sqrt{\frac{\sin(p-a)\sin(p-b)\sin(p-c)}{\sin p}}.$$

Приведенные формулы позволяют по любым трем элементам произвольного сферического треугольника определить его остальные элементы, то есть решить треугольник.

Сферический треугольник называют *прямоугольным*, если один из его углов равен 90°. Пусть прямым является угол A, тогда сторона a – гипотенуза

треугольника, стороны b и c – его катеты. Для прямоугольного сферического треугольника зависимости между его элементами выражаются более простыми формулами:

$$\sin b = \sin a \sin B; \quad (7.147)$$

$$\cos a = \cos b \cos c; \quad (7.148)$$

$$\sin a \cos B = \cos b \sin c. \quad (7.149)$$

Формулы упрощаются и тогда, когда одна из сторон, например a, равна 90°:

$$\cos A = \cos B \cos C; \quad (7.150)$$

$$\sin A \cos b = \cos B \sin C. \quad (7.151)$$

В случае узких сферических треугольников, то есть треугольников, у которых одна из сторон, например a, намного меньше любой другой, могут применяться приближенные формулы, в которых косинус малого угла заменяется единицей, а его синус заменяется величиной самого угла. Так, для решения треугольника, изображенного на рис. 7.21, могут использоваться, в частности, формулы:

$$a\cos B = (c - b);$$

$$A \sin b = a \sin B.$$

Для решения малых сферических треугольников могут использоваться формулы тригонометрии на плоскости.

Величина сферического избытка в произвольном треугольнике на сфере может быть вычислена по формуле

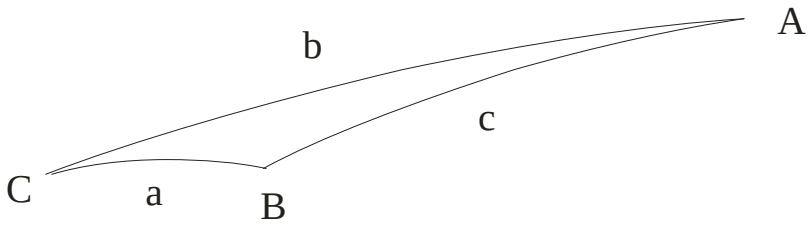


Рис. 7.21. Узкий треугольник

$$\sin\frac{\varepsilon}{2} = \frac{\sin\frac{a}{2}\sin\frac{b}{2}\sin C}{\cos\frac{c}{2}}, \quad (7.152)$$

в которой a, b и c – стороны треугольника, выраженные в угловой мере; C – угол, лежащий напротив стороны c, а также по другим подобным формулам.

В редких случаях значение сферического избытка может быть получено непосредственно, без использования формулы (7.152). Так, рассмотрим сферический треугольник, образованный дугой экватора с разностью долгот 90° и двумя меридианами, соединяющими концы этой дуги с северным полюсом. Все углы (и все стороны) такого треугольника равны 90° и его сферический избыток также составляет 90°.

7.7. Элементы большого круга

Положение большого круга в целом однозначно характеризуется долготой L точки его пересечения с экватором Q и углом между ним и экватором (рис. 7.22). Заимствуя некоторые термины из сферической астрономии, точку пересечения большого круга с экватором, которую будем считать *начальной*

точкой дуги большого круга, можно назвать *восходящим узлом*, а угол между экватором и большим кругом – *наклонением* большого круга. При движении точки *M* по дуге большого круга изменяются ее долгота, широта, длина дуги от экватора и другие параметры. Чтобы вывести зависимости между элементами дуги большого круга, которые могут оказаться полезными при решении тех или иных задач на сфере, рассмотрим рис. 7.23.

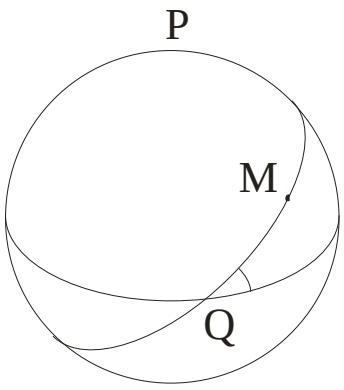


Рис. 7.22. Дуга большого круга

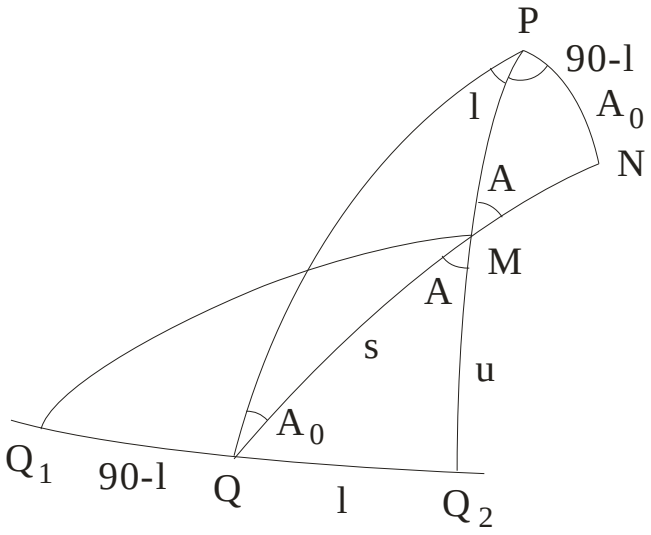


Рис. 7.23. Дополнительные треугольники

На рис. 7.23 *QMN* – дуга большого круга, по которому перемещается точка *M*; *Q* – начальная точка дуги большого круга; *N* – *вертексная точка* – точка с наибольшим значением широты. Следовательно, величина дуги *QMN* составляет 90°. Точка *P* – полюс сферы. Дуга Q_2MP – меридиан, проходящий через точку *M*, его дуга также равна 90°, а Q_2 – точка пересечения меридиана точки *M* с экватором. Q_1M – дуга *первого вертикала* – большого круга, проходящего через точку *M* и перпендикулярного меридиану этой точки; Q_1 – точка пересечения первого вертикала с экватором. Длина дуги Q_1M равна 90°. Дуга *QP* = 90° – меридиан, проходящий через начальную точку *Q* дуги большого круга.

На данном рисунке приняты следующие обозначения для пяти элементов большого круга:

u – широта точки *M*;

l – разность долгот точки *M* и точки *Q*;

s – дуга между точками *Q* и *M*;

A – азимут дуги большого круга в точке *M* – угол, отсчитываемый по часовой стрелке от северного направления меридиана, проходящего через *M*, до направления касательной к дуге большого круга в точке *M*;

A_0 – *начальный азимут* дуги большого круга – является дополнением наклонения до 90°.

PMQ – *полярный треугольник* – треугольник, одна из вершин которого является полюсом; QMQ_2 – прямоугольный треугольник; *PMN* – полярный треугольник с прямым углом в вершине *N*; Q_1MQ – треугольник, в котором угол при вершине *M* является дополнением азимута до 90°. Таким образом, в каждом из этих четырех треугольников один из элементов (сторона или угол) равен 90°. Применяя только формулу синусов, можно получить, например, следующий набор соотношений между элементами большого круга:

– из треугольника *PMQ*

$$\cos u \sin A = \sin A_0; \quad (7.153)$$

$$\sin s \sin \mathrm{A} = \sin l; \quad (7.154)$$

– из треугольника QMQ_2

$$\sin s \cos A_0 = \sin u; \quad (7.155)$$

$$\sin l \cos A_0 = \sin u \sin A; \quad (7.156)$$

– из треугольника *PMN*

$$\cos l \cos u = \cos s; \quad (7.157)$$

$$\cos l \sin A_0 = \cos s \sin \mathrm{A}; \quad (7.158)$$

– из треугольника Q_1MQ

$$\cos A_0 \cos \mathrm{l} = \cos A; \quad (7.159)$$

$$\cos l \sin u = \sin s \cos A. \quad (7.160)$$

На основе данных формул можно получить множество других для определения элементов дуги большого круга по двум известным. В частности, разделив равенство (7.156) на равенство (7.153), находим простое и удобное выражение зависимости между координатами точки большого круга:

$$\mathrm{tg} u = \mathrm{ctg} A_0 \sin l, \quad (7.161)$$

в котором A_0 является константой для фиксированного большого круга, а разность долгот l отсчитывается от его начальной точки. Изменяя значение константы A_0, можно получить все множество больших кругов с одной и той же начальной точкой.

Таким образом, число элементов дуги большого круга на сфере, отсчитываемой от точки ее пересечения с экватором, равно пяти. Этими элементами являются: начальный азимут A_0, широта *u* точки *M*, перемещающейся по дуге, разность долгот *l* начальной точки дуги на экваторе и точки *M*, дуга *s* от начальной точки до точки *M* и азимут *A* дуги в точке *M*. В табл. 7.3. приводится сводка коротких формул, позволяющих по любым двум известным элементам дуги большого круга *непосредственно* вычислить ее остальные элементы. Число таких формул равно числу сочетаний по три из пяти элементов

$$C_n^m = \frac{m!}{n!(m-n)!} = \frac{5!}{3!\,2!} = 10.$$

Таблица 7.3. Сводка формул для дуги большого круга

№ п/п	Формула
1	$\sin A_0 = \cos u \sin A$
2	$\mathrm{tg} A_0 = \mathrm{tg} A \cos s$
3	$\cos A = \cos A_0 \cos l$
4	$\mathrm{tg} l = \mathrm{tg} s \sin A_0$
5	$\mathrm{tg} l = \mathrm{tg} A \sin u$
6	$\sin l = \sin s \sin A$
7	$\cos s = \cos l \cos u$
8	$\sin u = \cos A_0 \sin s$
9	$\mathrm{tg} u = \mathrm{ctg} A_0 \sin l$
10	$\mathrm{tg} u = \mathrm{tg} s \cos A$

При пользовании данной сводкой необходимо найти формулу с нужным сочетанием элементов, а уже из нее получить значение определяемого элемента.

7.8. Главные геодезические задачи на сфере

Главными геодезическими задачами называют прямую геодезическую задачу и обратную геодезическую задачу. *Прямая геодезическая задача на сфере* заключается в том, чтобы по заданным координатам одной точки (широте u_1 и долготе L_1), азимуту A_1 и длине дуги s большого круга, выраженной в угловой мере, найти координаты второй точки и обратный азимут A_2 дуги большого круга, соединяющего эти точки (рис. 7.24). Если дуга s большого круга представлена в линейной мере как d, то ее переводят в угловую меру по формуле $s = d / R$, где R – радиус сферы.

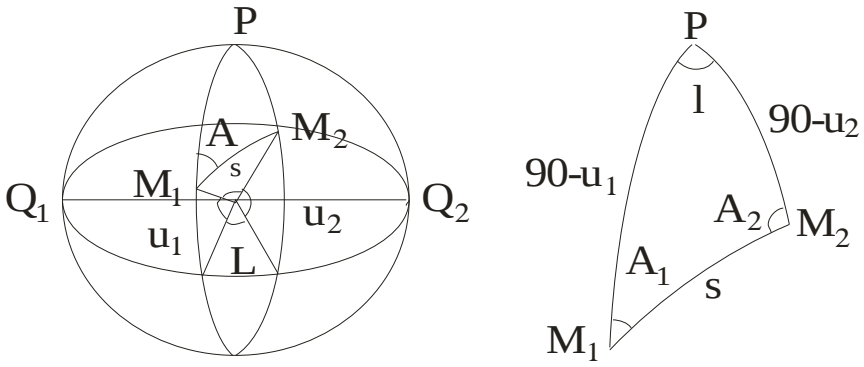


Рис. 7.24. Главная геодезическая задача

Решение прямой задачи можно получить из решения *полярного треугольника* (на рис. 7.24 справа). Применяя формулу косинуса стороны, находим

$$\sin u_2 = \sin u_1 \cos s + \cos u_1 \sin c \cos A_1 . \qquad (7.162)$$

Теперь можно дважды воспользоваться формулой синусов и получить:

$$\sin l = \frac{\sin s \sin A_1}{\cos u_2}, \qquad (7.163)$$

где l – разность долгот исходной и определяемой точек;

$$\sin A_2 = \frac{\cos u_1 \sin A_1}{\cos u_2}. \qquad (7.164)$$

Долгота определяемой точки равна

$$L_2 = L_1 + l. \qquad (7.165)$$

Обратная геодезическая задача на сфере заключается в нахождении по координатам двух точек u_1, L_1, u_2 и L_2 дуги s большого круга, проходящего через заданные точки, а также прямого A_1 и обратного A_2 азимутов.

Определив разность долгот

$$l = L_2 - L_1,$$

по формуле косинуса стороны находим

$$\cos s = \sin u_1 \sin u_2 + \cos u_1 \cos u_2 \cos l, \qquad (7.166)$$

а по формуле синусов:

$$\sin A_1 = \frac{\sin l}{\sin s}\cos u_2; \qquad (7.167)$$

$$\sin A_2 = \frac{\sin l}{\sin s}\cos u_1. \qquad (7.168)$$

7.9. Эллипсоид

Эллипсоидом (*трехосным эллипсоидом*) называют замкнутую центральную поверхность второго порядка. Если начало прямоугольной системы пространственных координат совпадает с *центром эллипсоида*, то его *каноническое уравнение* в такой системе координат имеет вид

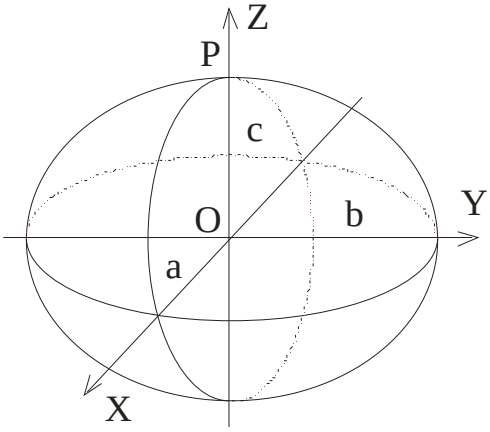
$$\frac{X^2}{a^2} + \frac{Y^2}{b^2} + \frac{Z^2}{c^2} = 1, \qquad (7.169)$$

где константы *a*, *b* и *c* – *полуоси* эллипсоида (рис. 7.25).

Сжатым эллипсоидом вращения, или *сфероидом*, называют эллипсоид, у которого две полуоси равны и каноническое уравнение которого имеет вид

$$\frac{X^2}{a^2} + \frac{Y^2}{a^2} + \frac{Z^2}{b^2} = 1, \qquad (7.170)$$

где *a* – большая и *b* – малая полуоси сфероида ($a > b$). Сфероид может быть получен вращением эллипса вокруг малой оси (рис. 7.26).



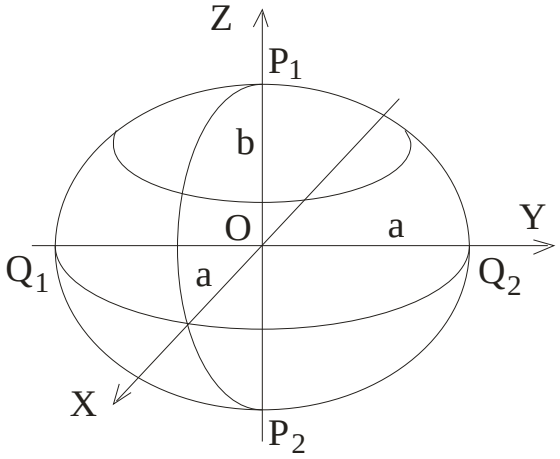


Рис. 7.25. Трехосный эллипсоид Рис. 7.26. Эллипсоид вращения

Хотя земная поверхность в целом ближе к поверхности трехосного эллипсоида, в высшей геодезии в качестве геометрической модели Земли принято использовать сфероид по двум причинам:

– его уравнение имеет более простой вид;

– по теоретическим соображениям: при остывании раскаленная жидкая масса планеты под действием сил взаимного притяжения и сил вращения при отсутствии возмущающих внешних сил должна была приобрести форму эллипсоида вращения.

Поэтому под земным эллипсоидом обычно понимают именно эллипсоид вращения.

Отрезок прямой, соединяющей любые две точки эллипсоида, называют хордой эллипсоида. Любая прямая, проходящая через центр эллипсоида вращения, пересекает его в двух точках, называемых диаметрально противоположными, а отрезок прямой, соединяющий эти точки, называют диаметром сфероида. Следовательно, большая и малая оси сфероида являются его диаметрами. Малая ось – его наименьший диаметр, а большая ось – наибольший диаметр и наибольшая хорда из всех возможных. Две диаметрально противоположные точки, соединяемые малой осью, называют северным и южным полюсами. Сечение сфероида любой плоскостью, содержащей малую ось, образует эллипс, называемый *меридианным*. Половину дуги эллипса, заключенную между двумя полюсами, называют *меридианом*.

На эллипсоиде, как и на сфере, вводятся два основных сечения:

– сечение плоскостью, проходящей через его центр и перпендикулярной малой оси *b*, называемое *экватором* эллипсоида вращения;

– некоторый меридиан, принимаемый за *начальный меридиан*.

Сечение эллипсоида плоскостью под произвольным углом Θ_0 к экватору образует эллипс, эксцентриситет которого зависит от угла Θ_0 . Эллипс, получаемый при $\Theta_0 = 90\ °$, то есть меридианный эллипс, характеризуется наибольшим значением эксцентриситета из всех возможных для данного эллипсоида. Сечение эллипсоида любой плоскостью, перпендикулярной малой

оси, то есть при $\Theta_0 = 0$ °, дает окружность (эллипс с наименьшим эксцентриситетом $e = 0$), называемую *параллелью*. Параллель обладает тем свойством, что для всех ее точек $Z = const$. Таким образом, экватор представляет собой параллель с наибольшим радиусом, для которой $Z = 0$, и делит сфероид на две симметричные части: *северный и южный полусфероиды*.

Выбор того или иного земного эллипсоида зависит от поставленных целей. Если определение параметров эллипсоида осуществляется при условии его максимальной близости к земной поверхности в целом, то такой эллипсоид называют *общим земным эллипсоидом*. Если параметры эллипсоида выбираются так, чтобы он наилучшим образом подходил для некоторого региона, например страны, то его называют *референц-эллипсоидом*.

Принятый в нашей стране эллипсоид Красовского характеризуется следующими параметрами:

большая полуось a = 6 378 245,000 00 м;

малая полуось b = 6 356 863,01877 м;

сжатие $\alpha = \frac{a-b}{a} = 0{,}003352329869$;

первый эксцентриситет $e^2 = \sin^2 \varepsilon = 0{,}006\,693\,421\,623$;

второй эксцентриситет $e'^2 = \mathrm{tg}^2 \varepsilon = 0{,}006\,738\,525\,415$,

по которым можно получить значения

ε = 4°41′34.093 905″ или ε = 0,081 904 878 551 радиан;

$\cos \varepsilon$ = 0,996 647 670 130;.

$k = \frac{1}{\cos \varepsilon} =$ 1,003 363 605 789.

7.10. Системы координат на эллипсоиде

При решении задач на поверхности эллипсоида в сфероидической геодезии используются следующие основные системы координат:

1) система прямоугольных пространственных координат;
2) система геоцентрических координат;
3) система геодезических координат;
4) система координат с приведенной широтой.

Наряду с ними используются система прямоугольных координат, отнесенных к плоскости меридиана точки на эллипсоиде, и некоторые другие системы.

*Система прямоугольных пространственных координа*т *XYZ* была неявным образом введена выше при определении канонических уравнений эллипсоида (7.169) и (7.170). Центр этой системы координат совпадает с центром эллипсоида, ось *OX* совпадает с линией пересечения плоскости экватора и плоскости начального меридиана, ось *OY* лежит в плоскости экватора, перпендикулярна к оси *OX* и образует с нею правую систему плоских координат в плоскости экватора; ось *OZ* совпадает с малой осью эллипсоида и направлена в сторону северного полюса.

Положение точки на эллипсоиде в остальных перечисленных основных системах координат определяется значениями ее долготы и широты. *Долгота L точки эллипсоида* в этих трех системах определяется одинаковым образом как угол между начальным меридианом и меридианом, проходящим через заданную точку. Таким образом, оставшиеся три системы координат различаются между собой определением широты точки, в связи с чем различают понятия геоцентрической, геодезической и приведенной широт.

Геоцентрическая широта φ точки на эллипсоиде есть угол между радиус-вектором ρ точки и плоскостью экватора. *Геодезическая широта B* точки эллипсоида определяется как угол между нормалью к поверхности эллипсоида в этой точке и плоскостью экватора. Трактовка приведенной широты *u* будет дана далее.

Длина радиус-вектора ρ точки эллипсоида в прямоугольной системе координат равна

$$\rho = \sqrt{X^2 + Y^2 + Z^2}\ ; \qquad (7.171)$$

ее долгота *L* наиболее просто определяется как

$$\mathrm{tg}L = \frac{Y}{X}, \qquad (7.172)$$

а геоцентрическая широта как

$$\mathrm{tg}\varphi = \frac{Z}{\sqrt{X^2 + Y^2}}. \qquad (7.173)$$

Обратный переход от геоцентрических координат к прямоугольным пространственным координатам может осуществляться по формулам:

$$\left.\begin{aligned} X &= \rho\cos\varphi\cos L \\ Y &= \rho\cos\varphi\sin L \\ Z &= \rho\sin\varphi \end{aligned}\right\}, \qquad (7.174)$$

где радиус-вектор

$$\rho = \frac{a}{\sqrt{\cos^2\varphi + k^2\sin^2\varphi}}. \qquad (7.175)$$

Для нахождения геодезической широты точки на эллипсоиде рассмотрим меридианный эллипс, получаемый в результате сечения эллипсоида плоскостью, проходящей через выбранную точку и содержащую малую полуось. Очевидно, что малая и большая полуоси этого эллипса будут соответственно равны малой и большой полуосям эллипсоида, а его уравнение имеет вид (7.2). Тангенс угла *B* между нормалью к эллипсу и производной $\frac{dy}{dx}$ связан соотношением

$$\mathrm{tg}B = -\frac{1}{\frac{dy}{dx}}. \qquad (7.176)$$

Продифференцировав (7.2) и подставив значение производной в (7.176), получим

$$\mathrm{tg}B = \frac{k^2 y}{x}, \qquad (7.177)$$

где, как и ранее, $k^2 = \dfrac{1}{\cos^2 \varepsilon} = \dfrac{1}{1-e^2}$.

Если в выражении (7.177) от системы плоских прямоугольных координат, отнесенных к плоскости меридианного эллипса, перейти к прямоугольным пространственным координатам, то оно примет вид

$$\mathrm{tg}B = \frac{k^2 Z}{\sqrt{X^2 + Y^2}}. \qquad (7.178)$$

Таким образом, мы получили зависимость между прямоугольными пространственными координатами точки на эллипсоиде и ее геодезической широтой *B*. С учетом равенства (7.173) находим наиболее простое соотношение между геодезической и геоцентрической широтами точки эллипсоида:

$$\mathrm{tg}B = k^2 \mathrm{tg}\varphi. \qquad (7.179)$$

Зависимость между приведенной широтой *u* и геоцентрической широтой φ выражается известной в сфероидической геодезии формулой

$$\mathrm{tg}u = k\,\mathrm{tg}\varphi, \qquad (7.180)$$

вывод которой будет дан далее. Из приведенных выражений можно получить множество других соотношений.

7.11. Кривые на поверхности эллипсоида

Произвольная кривая на эллипсоиде может быть представлена своими уравнениями вида $f_1(L,B)$, $f_2(L,\varphi)$ или $f_3(L,u)$ либо парой параметрических уравнений: $L = f_1(t)$, $B = f_2(t)$. В качестве параметра обычно выбирается длина кривой. Наиболее простыми линиями на эллипсоиде вращения являются его сечения плоскостью, или *плоские сечения*. Если секущая плоскость проходит через центр эллипсоида, то плоское сечение называют *центральным*. Если плоское сечение содержит нормаль к поверхности эллипсоида в некоторой фиксированной точке *m*, то такое плоское сечение называют *нормальным сечением* в этой точке. Очевидно, что через нормаль к поверхности эллипсоида в любой его точке можно провести бесконечное множество нормальных сечений.

Положение нормального сечения в заданной точке *m* характеризуется значением его азимута в этой точке. Кривизна нормального сечения в некоторой фиксированной точке *m* является функцией его азимута. При изменении азимута кривизна нормального сечения изменяется непрерывным образом в некотором диапазоне значений. В точке *m* на поверхности эллипсоида существуют два взаимно перпендикулярных направления, кривизна которых принимает минимальное и максимальное значения. Исключение представляют

полюсы эллипсоида вращения, являющиеся *омбилическими точками*, в которых кривизна по любому направлению является постоянной величиной. Нормальные сечения в точке m с минимальным и максимальным значениями кривизны называются *главными нормальными сечениями*, а их радиусы кривизны – *главными радиусами кривизны*. Одно из таких сечений совпадает с плоскостью меридиана, другое – с плоскостью *первого вертикала*, проходящей через нормаль к эллипсоиду и перпендикулярной к плоскости меридиана (рис. 7.27). Радиус кривизны меридиана обычно обозначают как M, а радиус кривизны первого вертикала – как N:

$$M = \frac{a(1-e^2)}{W^3}; \qquad (7.181)$$

$$N = \frac{a}{V}. \qquad (7.182)$$

Радиус кривизны нормального сечения на полюсе, называемый полярным радиусом, есть величина

$$c = \frac{a}{\sqrt{1-e^2}} = ka.$$

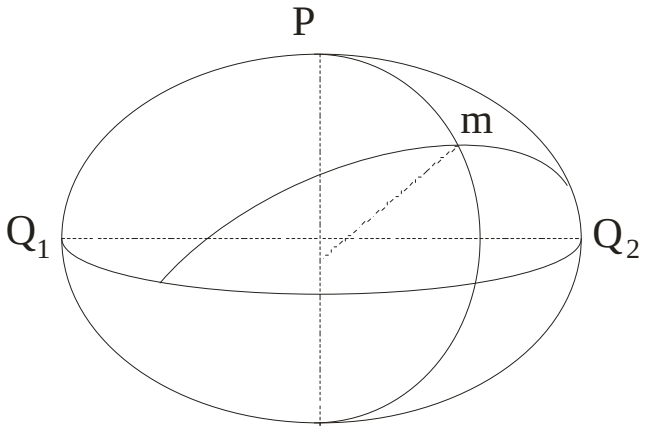


Рис. 7.27. Главные нормальные сечения

Если на поверхности эллипсоида взять две точки, то в каждой из них можно провести нормальное сечение таким образом, что оно будет проходить и через другую точку; такие сечения называют взаимно обратными сечениями, при этом одно из них называют прямым, а другое – обратным (рис. 7.28). В общем случае прямое и обратное сечения, проходящие через пару точек на эллипсоиде, не совпадают: если точка M_2 лежит севернее точки M_1, то ее прямое сечение будет проходить севернее обратного, являющегося прямым в M_1. Угол Δ между прямым и обратным нормальными сечениями может быть вычислен [9, с. 362] по формуле

$$\Delta = \frac{e^2 s^2}{4N_m^2}\cos^2 B_m \sin 2\alpha, \qquad (7.183)$$

где B_m – средняя геодезическая широта; s – расстояние между точками; α – азимут; N_m – радиус кривизны первого вертикала в точке с широтой B_m.

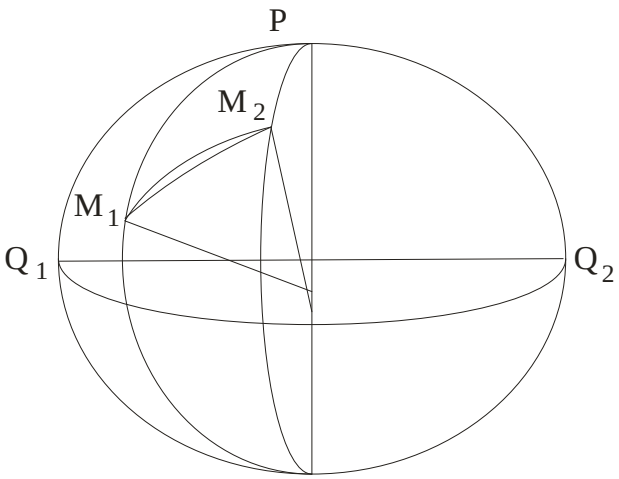


Рис. 7.28. Прямое и обратное сечения

Нормальные сечения представляют интерес в связи с измерением горизонтальных углов на поверхности эллипсоида. Если игнорировать уклонения отвесных линий, то нормаль к поверхности эллипсоида в каждой точке совпадает с отвесной линией в этой точке. Угломерный прибор (теодолит) устанавливается так, что его вертикальная ось совпадает с отвесной линией. Поэтому измеренные горизонтальные углы на поверхности эллипсоида представляют собой углы между прямыми сечениями. Этот факт имеет неприятное следствие: геометрические фигуры на поверхности эллипсоида оказываются незамкнутыми. На рис. 7.29 дан пример треугольника, в котором, по предположению, измерены три его угла, но в действительности это углы между прямыми сечениями.

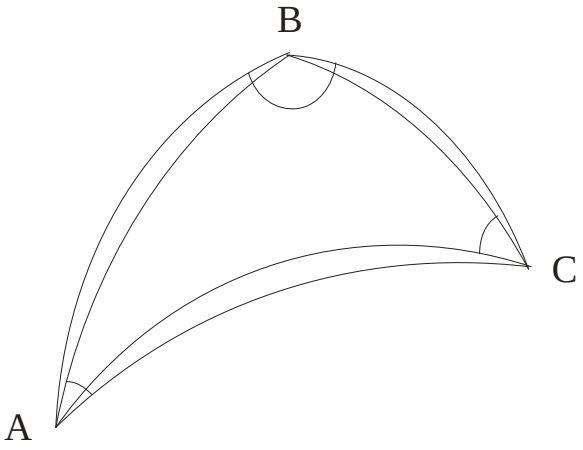


Рис. 7.29. Измеряемые углы

На плоскости кратчайшей линией между двумя точками является прямая, на сфере – дуга большого круга, а на эллипсоиде – кривая, называемая геодезической линией. В сфероидической геодезии использование геодезических линий позволяет устранить неопределенность в построении геометрических фигур на поверхности эллипсоида и достичь однозначности при решении задач [4].

Геодезическая линия на поверхности – это кривая, в любой точке которой соприкасающаяся плоскость проходит через нормаль к поверхности в этой точке, или иначе – кривая, в каждой точке которой главная нормаль к кривой совпадает с нормалью к поверхности. Геодезическая линия на эллипсоиде вращения не является плоской линией: она имеет не только кривизну, но и кручение.

Дифференциальное уравнение геодезической линии может быть записано как

$$\frac{d\alpha}{d\sigma} = \frac{\mathrm{tg}B}{N}\sin\alpha \,, \qquad (7.184)$$

где $d\alpha$ и $d\sigma$ – соответственно бесконечно малые приращения азимута и длины геодезической линии; α – ее азимут в текущей точке; N – радиус кривизны первого вертикала. Соотношение (7.184) определяет изменение

азимута при движении точки вдоль геодезической линии. Из него можно вывести дифференциальное уравнение в другой форме:

$$\frac{d\alpha}{dl} = \sin B, \qquad (7.185)$$

выражающее изменение азимута при изменении долготы точки на геодезической линии. В результате его интегрирования получают уравнение геодезической линии

$$r \sin \alpha = const, \qquad (7.186)$$

где r – радиус параллели, пересекаемой геодезической линией; α – азимут геодезической линии в точке пересечения. Из него следует, что произведение радиуса параллели на синус азимута геодезической линии есть постоянная величина. Уравнение (7.186) было получено Клеро, в связи с чем носит его имя, и определяет поведение геодезической линии на любой поверхности вращения.

Представить положение геодезической линии в целом на эллипсоиде с небольшим эксцентриситетом достаточно трудно. Это проще сделать, если рассматривать эллипсоид с большим значением эксцентриситета. На рис. 7.30 показана геодезическая линия, проходящая между точкой Q_1 на экваторе и точкой M, расположенной близко к экватору и далеко от Q_1.

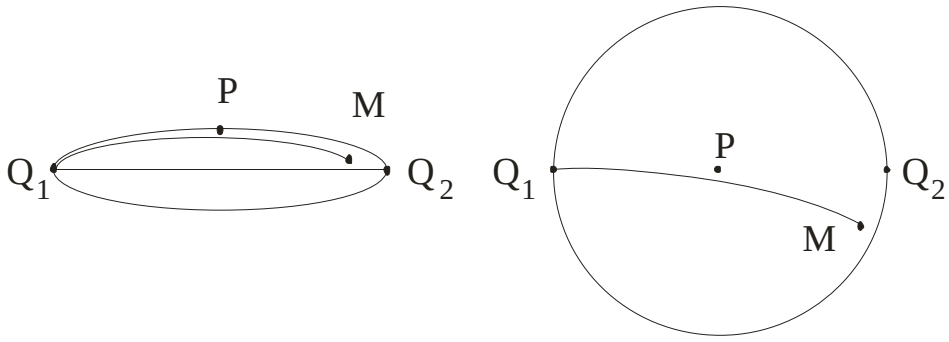


Рис. 7.30. Геодезическая линия на сфероиде

Из уравнения (7.186) с учетом $r = a \cos u$ можно получить выражение

$$\cos u \sin \alpha = \sin \alpha_0, \qquad (7.187)$$

где α_0 – азимут геодезической линии в точке ее пересечения с экватором, называемый начальным. Иногда равенство (7.187) записывают как

$$\cos u \sin \alpha = \cos u_0, \qquad (7.188)$$

где u_0 – приведенная широта точки геодезической линии, наиболее удаленной от экватора (вертексной точки).

Длина дуги геодезической линии может быть найдена путем интегрирования дифференциального уравнения

$$d\sigma = a\sqrt{1 - e^2 \cos^2 u}\, du, \qquad (7.189)$$

где $d\sigma$ – бесконечно малый элемент геодезической линии на эллипсоиде; a – большая полуось эллипсоида; u – приведенная широта точки.

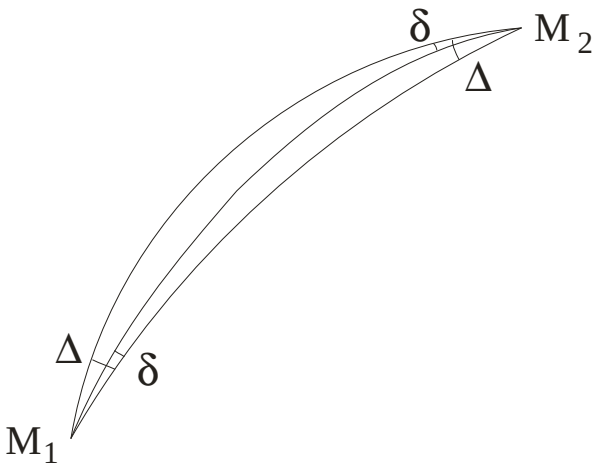


Рис. 7.31. Положение геодезической линии

Геодезическая линия, соединяющая две точки, проходит между их взаимными нормальными сечениями. Положение геодезической линии по отношению к взаимным нормальным сечениям показано на рис. 7.31. Угол δ между геодезической линией и прямым нормальным сечением составляет одну треть угла Δ между прямым и обратным сечениями:

$$\delta = \frac{\Delta}{3}.$$

Разность длин дуг нормального сечения и геодезической линии составляет величину

$$s - s_\Gamma = \frac{e^4 s^5}{360 N^4} \cos^4 B_m \sin^2 2\alpha \quad . \qquad (7.190)$$

7.12. Линейное отображение

Рассмотрение эллипса с проективной точки зрения позволило дать простую и ясную интерпретацию его параметров и сделать вывод некоторых формул сфероидической геодезии более очевидным и менее сложным. Выше на основе такого подхода было получено достаточно хорошее приближение эллиптического интеграла, названное квазиэллиптическим интегралом. Полученные результаты дают основания говорить о продуктивности рассмотрения некоторых вопросов сфероидической геодезии в проективно-геометрическом аспекте.

Поэтому представляется логичным, во-первых, распространить данный подход на более общий случай – трехмерное пространство, и, во-вторых, использовать при этом аппарат более общей теории – линейной алгебры. Ниже рассматривается линейное отображение одного евклидова пространства на другое, на основе которого далее будет дано решение главных задач сфероидической геодезии.

Отображение радиус-векторов. Пусть E – трехмерное евклидово пространство с системой прямоугольных декартовых координат oxyz, а F – такое же пространство с системой прямоугольных декартовых координат OXYZ. Пусть также задано отображение пространства E в пространство F:

$$\left.\begin{aligned} X &= kx \\ Y &= ky \\ Z &= z \end{aligned}\right\}, \qquad (7.191)$$

где k >1 – некоторая константа. Точке p (x, y, z) пространства E в пространстве F будет соответствовать точка P (X, Y, Z), которую принято называть образом точки p. Точку p при этом называют прообразом точки P.

В матричной форме отображение (7.191) можно записать как

$$\vec{R} = T \cdot \vec{r}, \qquad (7.192)$$

где $\vec{r}$ – радиус-вектор произвольной точки в пространстве E:

$$\vec{r} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}, \qquad (7.193)$$

$\vec{R}$ – образ радиус-вектора $\vec{r}$ в пространстве F:

$$\vec{R} = \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}, \qquad (7.194)$$

T – оператор отображения пространства E на пространство F

$$T = \begin{pmatrix} k & 0 & 0 \\ 0 & k & 0 \\ 0 & 0 & 1 \end{pmatrix}. \qquad (7.195)$$

Так как любая матрица вида

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \qquad (7.196)$$

задает в трехмерном евклидовом пространстве линейный оператор, то и оператор T является линейным. По определению линейный оператор обладает свойством

$$A(\lambda_1 \vec{x}_1 + \lambda_2 \vec{x}_2) = \lambda_1 A(\vec{x}_1) + \lambda_2 A(\vec{x}_2),$$

где λ – произвольные вещественные коэффициенты, а $\vec{x}$ – векторы. Оператор T является очень простым оператором, проще может быть только оператор подобия, получаемый из T заменой 1 на ту же константу k.

Отображение произвольных векторов. Произвольный вектор $\vec{v}$ в пространстве E может быть представлен как разность двух радиус-векторов

$$\vec{v} = \vec{r}_2 - \vec{r}_1, \qquad (7.197)$$

где $\vec{r}_1$ и $\vec{r}_2$ – радиус-векторы соответственно начальной и конечной точек вектора $\vec{v}$. В пространстве F образом вектора $\vec{v}$ является вектор $\vec{V}$, который также можно представить в виде разности соответствующих радиус-векторов

$$\vec{V} = \vec{R}_2 - \vec{R}_1.$$

Так как для любого радиус-вектора имеет место равенство

$$\vec{R} = T\,\vec{r},$$

то

$$\vec{V} = T\,\vec{r}_2 - T\,\vec{r}_1.$$

На основании свойства линейности оператора T последнее выражение можно представить как

$$\vec{V} = T(\vec{r}_2 - \vec{r}_1).$$

С учетом (7.197) получаем общую формулу

$$\vec{V} = T\,\vec{v}, \qquad (7.198)$$

описывающую отображение произвольных векторов пространства E на множество векторов пространства F.

Отображение плоскостей и прямых. Особенностью линейного оператора является то, что при его применении плоскости отображаются в плоскости, а прямые – в прямые. Если в пространстве E плоскость задана общим уравнением

$$ax + by + cz + d = 0,$$

то в пространстве F ее образом будет плоскость, представляемая общим уравнением

$$AX + BY + CZ + D = 0,$$

коэффициенты которого связаны с коэффициентами предыдущего уравнения равенствами

$$A = \frac{a}{k}; \quad B = \frac{b}{k}; \quad C = c; \quad D = d.$$

Проекции нормального вектора плоскости на оси координат равны соответствующим коэффициентам общего уравнения плоскости. Нормальные векторы $\vec{n}$ и $\vec{N}$ плоскостей в E и F могут быть представлены выражениями

$$\vec{n} = a\,\vec{e}_x + b\,\vec{e}_y + c\,\vec{e}_z$$

и

$$\vec{N} = A\,\vec{f}_X + B\,\vec{f}_Y + C\,\vec{f}_Z,$$

где $e = \{\vec{e}_x, \vec{e}_y, \vec{e}_z\}$ и $f = \{\vec{f}_X, \vec{f}_Y, \vec{f}_Z\}$ образуют ортонормированные базисы соответственно в пространствах E и F.

Прямые задаются в каждом пространстве уравнениями двух плоскостей.

Отображение длин векторов. Произвольный вектор $\vec{v}$ может быть представлен как сумма своих компонент

$$\vec{v} = \vec{v}_{xy} + \vec{v}_z,$$

где $\vec{v}_{xy}$ – векторная проекция вектора $\vec{v}$ на плоскость oxy, а $\vec{v}_z$ – его векторная проекция на ось oz. Значения модулей компонент определяются выражениями

$$\left.\begin{aligned} v_{xy} &= v\cos u \\ v_z &= v\sin u \end{aligned}\right\},$$

где $v,\ v_{xy},\ v_z$ – длины соответствующих векторов, u – угол между вектором $\vec{v}$ и плоскостью oxy. Аналогичным образом можно представить и вектор $\vec{V}$:

$$\vec{V} = \vec{V}_{XY} + \vec{V}_Z,$$

компоненты которого выражаются через компоненты вектора $\vec{v}$:

$$\left.\begin{aligned} \vec{V}_{XY} &= k\,\vec{v}_{xy} \\ \vec{V}_Z &= \vec{v}_z \end{aligned}\right\},$$

откуда следует, что длины векторов, параллельных оси oz, не искажаются, а масштаб отображения векторов, ортогональных оси oz, равен k. Масштаб μ отображения произвольного вектора $\vec{v}$ будет равен

$$\mu = \frac{V}{v} = \frac{\sqrt{k^2 v^2 \cos^2 u + v^2 \sin^2 u}}{v} = \sqrt{k^2 \cos^2 u + \sin^2 u}. \quad (7.199)$$

Полученное выражение определяет масштаб отображения вдоль любой прямой, расположенной под углом u к плоскости oxy в пространстве E. Для вычисления μ может использоваться также любое выражение из числа приведенных выше.

Отображение углов. Значение угла L в пространстве E между осью ox и проекцией вектора $\vec{v}$ на плоскость oxy может быть получено из равенства

$$\mathrm{tg}L = \frac{v_y}{v_x}, \quad (7.200)$$

где v_x и v_y – скалярные проекции вектора на оси координат ox и oy. Образом угла L в пространстве F будет являться угол Λ:

$$\mathrm{tg}\Lambda = \frac{V_Y}{V_X} = \frac{kv_y}{kv_x} = \mathrm{tg}L, \quad (7.201)$$

что означает сохранение ориентации векторов в плоскости, ортогональной оси oz.

Отображением вектора, лежащего в плоскости, параллельной оси oz пространства E, и образующего с плоскостью oxy угол u

$$\mathrm{tg}u = \frac{v_z}{\sqrt{{v_x}^2 + {v_y}^2}} \quad (7.202)$$

в пространстве F будет угол φ

$$\mathrm{tg}\varphi = \frac{V_Z}{V_{XY}} = \frac{v_z}{kv_{xy}} = \frac{1}{k}\mathrm{tg}u. \quad (7.203)$$

Вычисление угла φ может также осуществляться по формулам:

$$\left.\begin{aligned}\sin\varphi &= \frac{\sin u}{\mu}\\ \cos\varphi &= \frac{k\cos u}{\mu}\end{aligned}\right\}. \qquad (7.204)$$

Из данных формул следует, что если вектор параллелен (или ортогонален) оси oz в пространстве E, то его образ в пространстве F также будет параллелен (ортогонален) оси OZ.

Отображение длин и направлений на произвольной плоскости. Пусть в пространстве E задана плоскость p, содержащая вектор градиента $\vec{g}$ и ортогональный к нему вектор $\vec{h}$, а также вектор

$$\vec{v} = \vec{g} + \vec{h},$$

образующий с вектором градиента угол A (рис. 7.32). В пространстве F им будут соответствовать плоскость P и векторы

$$\vec{V} = \vec{G} + \vec{H},$$

а углу A – угол α. Векторы $\vec{G}$ и $\vec{H}$ будут ортогональны, поэтому масштаб m по направлению α на плоскости P (т. е. вдоль вектора $\vec{V}$) может быть вычислен по формуле

$$m = \frac{V}{v} = \frac{\sqrt{G^2 + H^2}}{v} = \frac{\sqrt{\mu^2 v^2 \cos^2 A + k^2 v^2 \sin^2 A}}{v} =$$

$$= \sqrt{\mu^2 \cos^2 A + k^2 \sin^2 A}, \qquad (7.205)$$

где масштаб μ по направлению градиента $\vec{G}$ может быть определен по формуле (7.199); масштаб вдоль вектора $\vec{H}$ равен k, так как вектор $\vec{h}$ ортогонален оси oz.

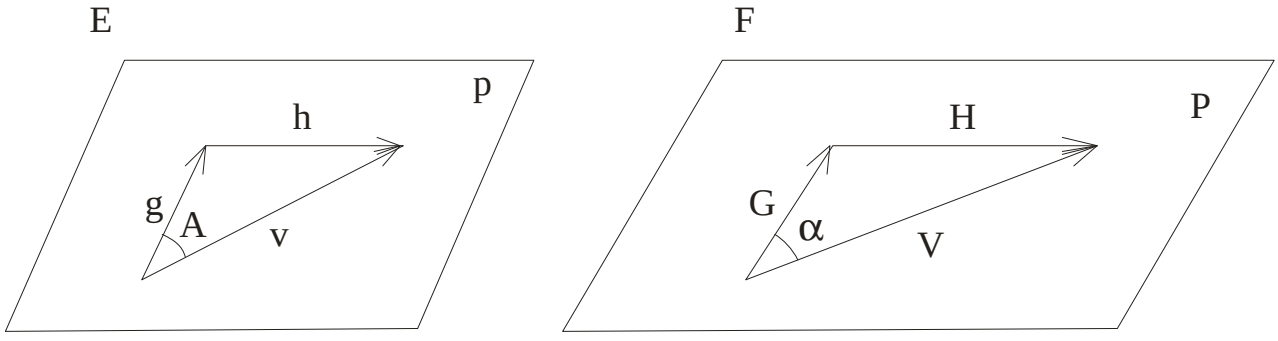


Рис. 7.32. Отображение векторов на плоскости

Угол α между вектором $\vec{G}$ и вектором $\vec{V}$ определяется соотношением

$$\operatorname{tg}\alpha = \frac{H}{G} = \frac{k\, v \sin A}{\mu\, v \cos A} = \frac{k}{\mu}\operatorname{tg}A, \qquad (7.206)$$

являющимся обобщением выражений (7.201) и (7.203).

Отображение сферы. Если в пространстве E задана сфера S с центром в начале координат и радиусом r

$$x^2 + y^2 + z^2 = r^2,$$

то в пространстве F ее образом будет эллипсоид

$$\frac{X^2}{a^2} + \frac{Y^2}{a^2} + \frac{Z^2}{b^2} = 1,$$

малая полуось которого $b = r$, большая полуось $a = kr$, а значения эксцентриситетов и константа k связаны соотношениями

$$e^2 = \frac{k^2 - 1}{k^2}; \quad e'^2 = k^2 - 1; \quad k^2 = \frac{1}{1 - e^2}. \qquad (7.207)$$

Системы сферических координат. Введем в пространстве E систему сферических координат {u, L, r}. Угол u между радиус-вектором r и плоскостью oxy будет приведенной широтой. Угол L – долгота, отсчитываемая от начального меридиана L = 0, лежащего в плоскости oxz. В пространстве F определим систему сферических координат {φ, Λ, ρ}. Угол φ между радиус-вектором ρ точки на эллипсоиде и плоскостью OXY будет ее геоцентрической широтой. Долготы Λ точек эллипсоида будем отсчитывать от плоскости OXZ, содержащей начальный меридиан Λ = 0. Разности долгот на сфере S и эллипсоиде будем обозначать соответственно как l и λ.

При отображении (7.192) и определении сферических координат вышеуказанным способом элементы φ, Λ и ρ в пространстве F будут образами соответственно элементов u, L и r пространства E.

Отображение полярных треугольников. Рассмотрим образованный дугами больших кругов на сфере в пространстве E треугольник pqm, в котором p – точка полюса, q – точка экватора, а m – произвольная точка. Его образом в пространстве F будет сфероидический треугольник PQM (рис. 7.33). Элементы сферического треугольника обозначены латинскими буквами, элементы сфероидического – греческими. Очевидно, что каждая дуга большого круга на сфере в пространстве E будет трансформироваться в плоскую кривую в пространстве F, получающуюся в результате пересечения поверхности эллипсоида с плоскостью, проходящей через начало координат.

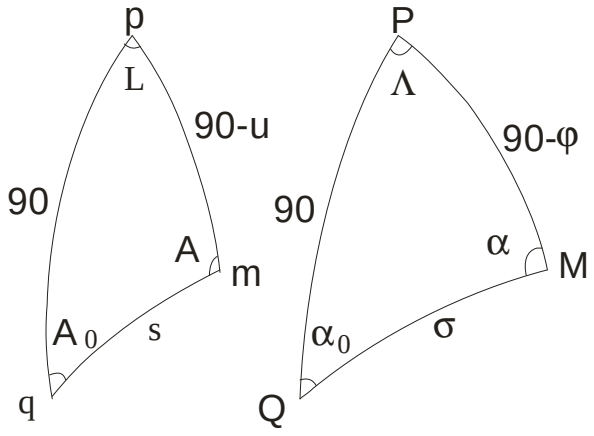


Рис. 7.33. Линейное отображение полярного треугольника

Рассмотрим связь элементов полярного сфероидического

треугольника с их прообразами из пространства E. Дуга меридиана PQ будет равна (в угловой мере)

$$PQ = pq = \frac{\pi}{2}. \qquad (7.208)$$

В силу соотношения (7.201) долгота точки на эллипсоиде будет равна долготе точки на сфере:

$$\Lambda = L. \qquad (7.209)$$

Начальный азимут плоской кривой на эллипсоиде, как можно установить с учетом (7.206), связан с начальным азимутом дуги большого круга на сфере соотношениями

$$\left.\begin{aligned} \operatorname{tg}\alpha_0 &= k\ \operatorname{tg}A_0 \\ \sin\alpha_0 &= \frac{k\sin A_0}{m_0} \\ \cos\alpha_0 &= \frac{\cos A_0}{m_0} \end{aligned}\right\}, \qquad (7.210)$$

где m_0 – масштаб на экваторе эллипсоида по азимуту α_0:

$$m_0 = \sqrt{k^2\sin^2 A_0 + \cos^2 A_0}. \qquad (7.211)$$

Геоцентрическая широта точки на эллипсоиде может быть выражена через приведенную широту в соответствии с (7.203) и (7.204)) как

$$\left.\begin{aligned} \operatorname{tg}\varphi &= \frac{1}{k}\operatorname{tg}u \\ \sin\varphi &= \frac{\sin u}{\mu} \\ \cos\varphi &= \frac{k\cos u}{\mu} \end{aligned}\right\}, \qquad (7.212)$$

где масштаб μ определяется выражением (7.199).

На основании (7.206) и (7.210) можно установить взаимосвязь между дугой большого круга s и дугой плоской кривой σ на эллипсоиде:

$$\left.\begin{aligned} \operatorname{tg}\sigma &= \frac{m_0}{k}\operatorname{tg}s \\ \sin\sigma &= \frac{m_0 \sin s}{\mu} \\ \cos\sigma &= \frac{k\cos s}{\mu} \end{aligned}\right\}. \qquad (7.213)$$

Наконец, используя (7.206), можно определить, что азимут плоской кривой на эллипсоиде будет связан с азимутом дуги большого круга на сфере равенствами

$$\left.\begin{aligned} \operatorname{tg}\alpha &= \frac{k}{\eta}\operatorname{tg}A \\ \sin\alpha &= \frac{k\sin A}{m} \\ \cos\alpha &= \frac{\eta\cos A}{m} \end{aligned}\right\}, \qquad (7.214)$$

где η – тангенциальный масштаб в точке с широтой φ по направлению меридиана

$$\eta = \sqrt{k^2 \sin^2 u + \cos^2 u}$$

может быть получен заменой $\sin u$ и $\cos u$ в выражении (7.199) на их кофункции, а m – масштаб по направлению α на эллипсоиде равен

$$m = \sqrt{k^2 \sin^2 A + \eta^2 \cos^2 A}.$$

7.13. Центральная проекция эллипсоида на сферу

Все величины в сферическом и сфероидическом треугольниках мы рассматриваем как угловые. Важная особенность полярного сфероидического треугольника PQM состоит в том, что между его четырьмя элементами выполняется соотношение

$$\frac{\cos\varphi}{\sin\alpha_0} = \frac{\sin\sigma}{\sin\lambda},$$

в чем нетрудно убедиться, если в формуле синусов

$$\frac{\cos u}{\sin A_0} = \frac{\sin s}{\sin l}$$

элементы сферического треугольника заменить их значениями в соответствии с (7.209)–(7.213). Этот факт становится не столь неожиданным, если заметить, что все участвующие в последних двух выражениях величины и

углы, ограничивающие дуги меридианов pq и PQ, являются центральными углами.

Досадным диссонансом выглядит зависимость (7.214) между азимутом центрального сечения эллипсоида и его прообразом – азимутом дуги большого круга на сфере. Любые две величины из четырех указанных выше и дуга меридиана PQ однозначно определяют полярный сфероидический треугольник. Если бы была возможность измерять прямо или косвенно начальный азимут α_0 (угол σ может быть вычислен по результатам измерения других величин), то мы вполне могли бы ограничиться использованием только этих величин и определять по ним широту φ и разность долгот λ из решения сфероидических треугольников (прямая задача), либо по φ и λ вычислять σ и α_0 (обратная задача). Но реальный мир устроен так, что из угловых величин на поверхности эллипсоида для измерений нам доступны только углы между направлениями и азимуты линий.

Поэтому сам собой возникает вопрос о возможности такого преобразования азимута на эллипсоиде, чтобы к нему была применима теорема синусов, и задачи на эллипсоиде можно было бы решать так же просто, как и на сфере. Такая возможность существует. Чтобы преобразовать азимут α требуемым образом, построим в пространстве F сферу S' с радиусом $R = b$ и центром в начале координат и спроектируем все точки эллипсоида на эту сферу по нормалям к ней, то есть построим центральную проекцию эллипсоида на сферу. При таком проектировании все величины полярного сфероидического треугольника, за исключением азимута, сохранят свои значения как центральные углы (рис. 7.34).

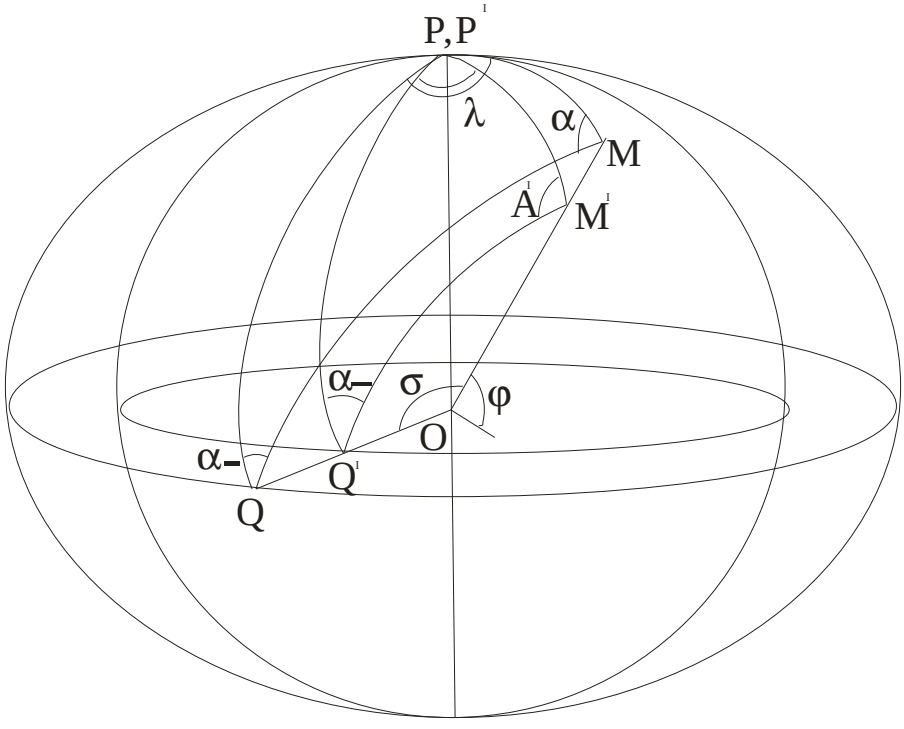


Рис. 7.34. Центральная проекция эллипсоида на сферу

Азимут α центрального сечения эллипсоида при этом трансформируется в азимут A' дуги большого круга на сфере S′. Значение последнего с учетом формулы 3 из табл. 7.3 может быть вычислено по формуле

$$\sin A' = \sqrt{\sin^2 \alpha_0 + \cos^2 \alpha_0 \sin^2 \lambda}, \qquad (7.215)$$

которая на основе (7.156) и (7.210) может быть приведена к более удобному виду

$$\sin A' = \frac{\mu}{m_0} \sin A. \qquad (7.216)$$

Соотношения (7.209)–(7.214), дополненные последним выражением, дают возможность установить следующую аналогию между элементами сферического и сфероидического треугольников pqm и $P'Q'M'$, представленную в табл. 7.4. В данной таблице для сравнения приведены формулы из табл. 7.3 и даны соотношения между элементами центрального сечения сфероида. Исключением является значение азимута A′, для которого в качестве рабочего названия может использоваться термин «приведенный азимут».

Таблица 7.4. Соотношения между элементами сферического и сфероидического треугольников

№ п/п	Элементы на сфере	Элементы на эллипсоиде
1	$\sin A_0 = \cos u \sin A$	$\sin \alpha_0 = \cos \varphi \sin A'$
2	$\mathrm{tg} A_0 = \mathrm{tg} A \cos s$	$\mathrm{tg} \alpha_0 = \mathrm{tg} \alpha \cos \sigma$
3	$\cos A = \cos A_0 \cos l$	$\cos A' = \cos \alpha_0 \cos \lambda$
4	$\mathrm{tg} l = \mathrm{tg} s \sin A_0$	$\mathrm{tg} \lambda = \mathrm{tg} \sigma \sin \alpha_0$
5	$\mathrm{tg} l = \mathrm{tg} A \sin u$	$\mathrm{tg} \lambda = \mathrm{tg} A' \sin \varphi$
6	$\sin l = \sin s \sin A$	$\sin \lambda = \sin \sigma \sin A'$
7	$\cos s = \cos l \cos u$	$\cos \sigma = \cos \lambda \cos \varphi$
8	$\sin u = \cos A_0 \sin s$	$\sin \varphi = \cos \alpha_0 \sin \sigma$
9	$\mathrm{tg} u = \mathrm{ctg} A_0 \sin l$	$\mathrm{tg} \varphi = \mathrm{ctg} \alpha_0 \sin \lambda$
10	$\mathrm{tg} u = \mathrm{tg} s \cos A$	$\mathrm{tg} \varphi = \mathrm{tg} \sigma \cos A'$

В данной сводке формул (которую можно продолжить) слева – соотношения между элементами сферического треугольника pqm, справа – между элементами сферического треугольника P′Q′M′, полученные заменой величин в левых формулах в соответствии с (7.209)–(7.214) и (7.216). Можно обратить внимание на выражение 9 в табл. 7.4, устанавливающее очень простую связь между координатами точки центрального сечения эллипсоида вращения.

7.14. Решение главных геодезических задач на эллипсоиде

Методы решения главных геодезических задач оказывают доминирующее влияние на точность и вычислительную эффективность определения координат.

Практическая значимость этой задачи для сфероидической геодезии подчеркивается не только названием, но и тем фактом, что ее решением занимались многие выдающиеся математики и геодезисты. Возможно, что число вариантов ее решения составляет до двух десятков. Наиболее известны из них методы, предложенные Ф.В. Бесселем, К.Ф. Гауссом, Ф.Н. Красовским, М.С. Молоденским, Ф.А. Слудским и некоторыми другими геодезистами.

Тем не менее, изменения в методах геодезических измерений настолько радикальны, что возникла необходимость осмыслить методы решения главных геодезических задач. Ниже излагается точный метод решения главных геодезических задач с применением центральных сечений эллипсоида.

Выше мы выполнили необходимую подготовительную работу и теперь можем перейти к непосредственному решению главных геодезических задач на эллипсоиде. Идея решения заключается в том, что если мы получим каким-либо способом центральные углы φ и λ на сфере S', то тем самым определим геоцентрическую широту и долготу точки на эллипсоиде.

Решение прямой задачи. В предлагаемой постановке прямая геодезическая задача на эллипсоиде сводится к тому, чтобы по заданным величинам: φ_1 – геоцентрической широте исходной точки 1, Λ_1 – ее долготе, α_{12} – азимуту дуги центрального сечения 1–2 в исходной точке и σ – центральному углу, ограничивающему дугу плоской кривой, вычислить геоцентрические координаты φ_2 и Λ_2 определяемой точки 2 и обратный азимут α_{21} дуги центрального сечения на эллипсоиде. Решение поставленной задачи выполняется следующим образом.

1. Переход от азимута α на эллипсоиде к азимуту A′ на сфере S′:

$$\operatorname{tg} A'_{12} = \frac{\operatorname{tg}\alpha_{12}}{\cos(B_1 - \varphi_1)}. \qquad (7.217)$$

Данная формула следует из того, что угол между касательной плоскостью в точке 1 на эллипсоиде и касательной плоскостью в точке на сфере, являющейся проекцией точки 1, равен $(B_1 - \varphi_1)$.

2. Нахождение геоцентрической широты по формуле косинуса стороны

$$\sin\varphi_2 = \sin\varphi_1 \cos\sigma + \cos\varphi_1 \sin\sigma \cos A'_{12}. \qquad (7.218)$$

3. Определение разности долгот по формуле синусов

$$\sin\lambda = \frac{\sin\sigma}{\cos\varphi_2} \sin A'_{12}. \qquad (7.219)$$

4. Вычисление долготы определяемой точки

$$A_2 = A_1 + \lambda. \qquad (7.220)$$

5. Определение обратного азимута на сфере по формуле синусов

$$\sin A'_{21} = \frac{\cos\varphi_1}{\cos\varphi_2} \sin A'_{12}. \qquad (7.221)$$

6. Переход от обратного азимута на сфере к обратному азимуту на эллипсоиде по формуле, аналогичной (7.217):

$$\operatorname{tg}\alpha_{21} = \cos(B_2 - \varphi_2)\operatorname{tg}A'_{21} \qquad (7.222)$$

или

$$\operatorname{tg}\alpha_{21} = \operatorname{tg}A'_{21}\frac{1+(k^2-1)\sin^2\varphi_2}{\sqrt{1+(k^4-1)\sin^2\varphi_2}}. \qquad (7.223)$$

Решение обратной задачи. В предлагаемой постановке обратная геодезическая задача на эллипсоиде состоит в том, чтобы по заданным геоцентрическим координатам φ_1, Λ_1 и φ_2, Λ_2 двух точек найти прямой α_{12} и обратный α_{21} азимуты дуги центрального сечения эллипсоида и ограничивающий ее центральный угол σ. Для решения задачи необходимо выполнить следующие действия.

1. Определение угла σ по формуле косинусов

$$\cos\sigma = \sin\varphi_1\sin\varphi_2 + \cos\varphi_1\cos\varphi_2\cos(\Lambda_2 - \Lambda_1). \qquad (7.224)$$

2. Вычисление прямого и обратного азимутов на сфере

$$\left.\begin{aligned}\sin A'_{12} &= \frac{\sin(\Lambda_2-\Lambda_1)}{\sin\sigma}\cos\varphi_2\\ \sin A'_{21} &= \frac{\sin(\Lambda_2-\Lambda_1)}{\sin\sigma}\cos\varphi_1\end{aligned}\right\}. \qquad (7.225)$$

3. Переход от азимутов на сфере к азимутам плоской кривой на эллипсоиде по формулам

$$\left.\begin{aligned}\operatorname{tg}\alpha_{12} &= \cos(B_1-\varphi_1)\operatorname{tg}A'_{12}\\ \operatorname{tg}\alpha_{21} &= \cos(B_2-\varphi_2)\operatorname{tg}A'_{21}\end{aligned}\right\} \qquad (7.226)$$

или

$$\left.\begin{aligned}\operatorname{tg}\alpha_{12} &= \operatorname{tg}A'_{12}\frac{1+(k^2-1)\sin^2\varphi_1}{\sqrt{1+(k^4-1)\sin^2\varphi_1}}\\ \operatorname{tg}\alpha_{21} &= \operatorname{tg}A'_{21}\frac{1+(k^2-1)\sin^2\varphi_2}{\sqrt{1+(k^4-1)\sin^2\varphi_2}}\end{aligned}\right\} \qquad (7.227)$$

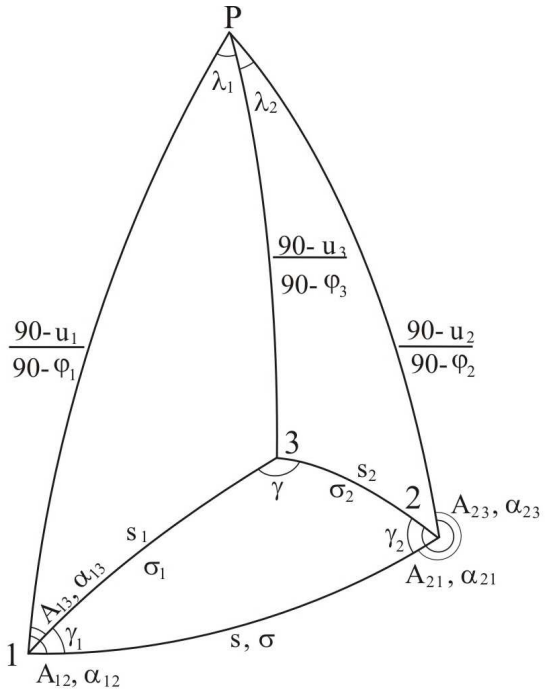


Рис. 7.35. Демонстрационный пример

Демонстрационный пример. Для иллюстрации предлагаемого метода решения главных геодезических задач ниже на макетных данных рассматривается вычисление азимутальной засечки на эллипсоиде

(рис. 7.35). Чтобы подчеркнуть возможности метода, эллипсоид преднамеренно выбран достаточно «плохой» – отношение его большой полуоси к малой равно 2, что соответствует эксцентриситету e2 = 3/4.

Первичными данными служат координаты трех точек 1, 2 и 3 на сфере S: $u_1 = 20°$, $L_1 = 10°$; $u_2 = 15°$, $L_2 = 80°$; $u_3 = 55°$, $L_3 = 40°$. Точки 1 и 2 считаются исходными, точка 3 – определяемой. По координатам точек на сфере S определяются координаты исходных точек на эллипсоиде и азимуты (также на эллипсоиде) с исходных точек на определяемую точку. Далее по этим величинам с целью демонстрации внутренней непротиворечивости дважды (из полярных треугольников P13 и P23) определяются геоцентрическая широта и долгота определяемой точки.

Для внешнего контроля геоцентрическая широта и долгота определяемой точки вычисляются непосредственно по ее координатам на сфере S с применением формул (7.212) и (7.209). В заключение вычисляется обратный азимут α_{31} по формулам (7.225), (7.226) и (7.227) и, для сравнения, по формуле (7.214). Результаты вычислений приводятся в табл. 7.5.

Таблица 7.5. Вычисление азимутальной засечки

Подготовка данных – определение координат точек и азимутов на эллипсоиде	
1) $\cos s_1 = \sin u_1 \sin u_3 + \cos u_1 \cos u_3 \cos(L_3 - L_1)$	0,746 941 673 572 354
2) $\sin A_{13} = \sin(L_3 - L_1)\cos u_3 / \sin s_1$	0,431 332 107 491 882
3) $\cos s_2 = \sin u_2 \sin u_3 + \cos u_2 \cos u_3 \cos(L_3 - L_2)$	0,636 425 509 428 080
4) $\sin A_{23} = \sin(L_3 - L_2)\cos u_3 / \sin s_2$	0,477 984 678 205 132
5) $\sin A_{013} = \cos u_1 \sin A_{13}$	0,405 319 598 518 156
6) $\sin A_{023} = \cos u_2 \sin A_{23}$	0,461 697 745 248 807
7) $\sin\alpha_{013} = k \sin A_{013} / \sqrt{k^2 \sin^2 A_{013} + \cos^2 A_{013}}$	0,663 466 856 227 746
8) $\sin\alpha_{023} = k \sin A_{023} / \sqrt{k^2 \sin A_{023} + \cos^2 A_{023}}$	0,721 161 905 091 940
9) $\sin\varphi_1 = \sin u_1 / \sqrt{\sin^2 u_1 + k^2 \cos^2 u_1}$	0,179 044 418 060 501
10) $\sin\varphi_2 = \sin u_2 / \sqrt{\sin^2 u_2 + k^2 \cos^2 u_2}$	0,132 788 176 096 108
11) $\sin\alpha_{13} = \sin\alpha_{013} / \cos\varphi_1$	0,674 363 908 422 641
12) $\sin\alpha_{23} = \sin\alpha_{023} / \cos\varphi_2$	0,727 605 257 344 666
Решение азимутальной засечки	
13) $\cos\sigma = \sin\varphi_1 \sin\varphi_2 + \cos\varphi_1 \cos\varphi_2 \cos(L_2 - L_1)$	0,357 288 580 672 199
14) $\sin\alpha_{12} = \sin(L_2 - L_1)\cos\varphi_2 / \sin\sigma$	0,997 191 710 446 353
15) $\sin\alpha_{21} = \sin(L_2 - L_1)\cos\varphi_1 / \sin\sigma$	0,989 843 701 639 624

16) $\sin\gamma_1 = \sin A'_{12}\cos A'_{13} - \cos A'_{12}\sin A'_{13}$	0,685 821 595 185 382
17) $\sin\gamma_2 = \sin A'_{23}\cos A'_{21} - \cos A'_{23}\sin A'_{21}$	0,575 592 561 784 983
18) $\cos\gamma = -\cos\gamma_1\cos\gamma_2 + \sin\gamma_1\sin\gamma_2\cos\sigma$	-0,454 082 918 687 637
19) $\sin\sigma_1 = \sin\sigma\sin\gamma_2 / \sin\gamma$	0,603 394 504 968 054
20) $\sin\sigma_2 = \sin\sigma\sin\gamma_1 / \sin\gamma$	0,718 947 758 185 017
21) $\sin\varphi_3 = \sin\varphi_1\cos\sigma_1 + \cos\varphi_1\sin\sigma_1\cos A'_{13}$	*0,581 124 101 764 717*
22) $\sin\lambda_1 = \sin\sigma_1\sin A'_{13} / \cos\varphi_3$	0,500 000 000 000 000
23) $\sin\lambda_2 = \sin\sigma_2\sin A'_{23} / \cos\varphi_3$	0,642 787 609 686 539
24) $\sin\Lambda_3 = \sin L_1\cos\lambda_1 + \cos L_1\sin\lambda_1$	*0,642 787 609 686 539*
25) $\sin A'_{31} = \sin\alpha_{013} / \cos\varphi_3$	0,815 255 179 903 676
26) $\mathrm{tg}B_3 = k^2 \cdot \mathrm{tg}\varphi_3$	2,856 296 013 384 227
27) $\mathrm{tg}\alpha_{31} = \mathrm{tg}A'_{31}\cos(B_3 - \varphi_3)$	*1,150 724 103 361 136*
Внутренний контроль	
28) $\sin\varphi_3 = \sin\varphi_2\cos\sigma_2 + \cos\varphi_2\sin\sigma_2\cos A'_{23}$	*0,581 124 101 764 717*
29) $\sin\Lambda_3 = \sin L_2\cos\lambda_2 - \cos L_2\sin\lambda_2$	*0,642 787 609 686 539*
30) $\mathrm{tg}\alpha_{31} = \mathrm{tg}A'_{31}\dfrac{1+(k^2-1)\cdot\sin^2\varphi_3}{\sqrt{1+(k^4-1)\cdot\sin^2\varphi_3}}$	*1,150 724 103 361 136*
Внешний контроль	
31) $\sin\varphi_3 = \sin u_3 / \sqrt{\sin^2 u_3 + k^2\cos^2 u_3}$	*0,581 124 101 764 717*
32) $\sin\Lambda_3 = \sin L_3$	*0,642 787 609 686 539*
33) $\eta_3 = \sqrt{k^2\cdot\sin^2 u_3 + \cos^2 u_3}$	1,735 808 231 052 181
34) $\sin A_{31} = \sin A_{013} / \cos u_3$	0,706 653 155 238 909
35) $\mathrm{tg}\alpha_{31} = \dfrac{k}{\eta_3}\mathrm{tg}A_{31}$	*1,150 724 103 361 136*

Вычислительные эксперименты проводились с эллипсоидами разной формы (до k =10) и с различными точками. Их результаты здесь не приводятся, но следует сказать, что при больших значениях k и очень плохих углах засечки начинают сказываться ошибки округления – неверными становятся 14-я и даже 13-я цифры после десятичной точки.

Задача внутренней оптимизации вычислений в данном примере не ставилась, его целью является демонстрация принципиальной возможности реализации метода.

7.15. Сравнение геодезических линий и центральных сечений

Как отмечалось выше, геодезические линии дают возможность получения на эллипсоиде вращения замкнутых геометрических фигур. Но их использование при решении задач на сфероиде не является безупречным и сопровождается определенными неудобствами. Еще Бессель отмечал: «Если говорить со всей строгостью, то не существует способов наблюдения углов треугольников, расположенных на поверхности Земли; стороны этих треугольников представляют собой геодезические линии, а наблюдаются только углы между вертикальными сечениями эллипсоида, проведенными от точки стояния инструмента до двух других вершин треугольника» [1, с. 25] и «Легко устранимо небольшое затруднение, заключающееся в том, что ни одной стороны треугольника нельзя измерить как геодезическую линию» [1, с. 29].

Таким образом, измеряемыми углами являются углы между прямыми нормальными сечениями. В триангуляции 1-го класса в измеренные направления вводятся поправки за переход к геодезическим линиям, в остальных случаях углы между геодезическими линиями принимаются равными углам между соответствующими нормальными сечениями на пункте. Разностью длин нормального сечения и геодезической линии при решении треугольников триангуляции пренебрегают.

Стороны геодезических построений ближе к прямым в пространстве, чем к геодезическим линиям на поверхности эллипсоида. Это особенно справедливо при электронных методах измерения расстояний. Если бы физическая поверхность Земли была гладкой и в нашем распоряжении имелись мерные ленты любой длины, то мы измеряли бы длину геодезических линий на эллипсоиде вращения. При использовании светодальномеров действительно измеряется длина геодезической линии, но не на эллипсоиде вращения, а *в пространстве*. Эта геодезическая линия имеет меньшую кривизну и близка к прямой между двумя точками пространства.

Как положительное свойство геодезических линий обычно отмечается то обстоятельство, что между двумя точками на сфероиде можно провести только одну геодезическую линию. Данное утверждение справедливо в большинстве случаев, но требует некоторых пояснений. Чтобы показать это, проведем мысленный эксперимент, воспользовавшись рис. 7.36.

Если бы нам пришлось определять площадь некоторой геометрической фигуры, одна из сторон которой совпадает с Q_1Q_2, то перед нами возникла бы нелегкая задача выбора одной геодезической линии из двух возможных. (Если бы дуга Q_1Q_2 являлась участком границы между двумя государствами, то это был бы повод для территориальных притязаний.) В данной ситуации разумным решением представляется отказ от геодезической линии и использование дуги экватора, то есть центрального сечения.

Предположим, что на экваторе расположены две точки, разность долгот l которых составляет, например, 80°. Очевидно, что между этими точками можно провести две симметричные геодезические линии. Дуга экватора между заданными точками также отвечает уравнениям геодезической линии (7.186) и

(7.187), и в каждой ее точке главная нормаль к ней совпадает с нормалью к эллипсоиду вращения. Но дуга экватора не является геодезической линией.

Если бы мы натянули упругую нить, совпадающую с геодезической линией между Q_1 и Q_2, и находились в точке Q_1, то, скорее всего, были бы несколько обескуражены тем фактом, что точку Q_2 мы видим в направлении дуги экватора, а геодезическая линия уходит от него под большим углом. Хотя причина такого отклонения понятна, в данном случае психологически трудно принять геодезическую линию в качестве основной кривой между указанными точками, поскольку мы привыкли измерять расстояния «по прямой». Естественно, что «прямое» направление – это то, в котором мы видим точку.

Продолжим наш мысленный эксперимент. Допустим, что на сфероиде нам удалось натянуть упругую нить между точками Q_1 (начальная точка) и Q_2 (конечная точка) так, что она совпадает с дугой экватора (рис. 7.37). Такое положение нити является положением неустойчивого равновесия. При малейшем смещении конечной точки упругой нити из точки Q_2 в точку M_1 и при отсутствии сил трения упругая нить скачком займет положение геодезической линии G_1. Если теперь непрерывным образом перемещать конец упругой нити в симметричную по отношению к M_1 точку M_2, находящуюся по другую сторону экватора, то упругая нить в конечном итоге совпадет с кривой G_3. Уравнение этой кривой также имеет вид (7.186) или (7.187), и в каждой ее точке главная нормаль совпадает с нормалью к поверхности, G_3 является геодезической линией, но не является кратчайшей кривой между этими точками. Кратчайшей линией между точками Q_1 и M_2 является геодезическая линия G_2.

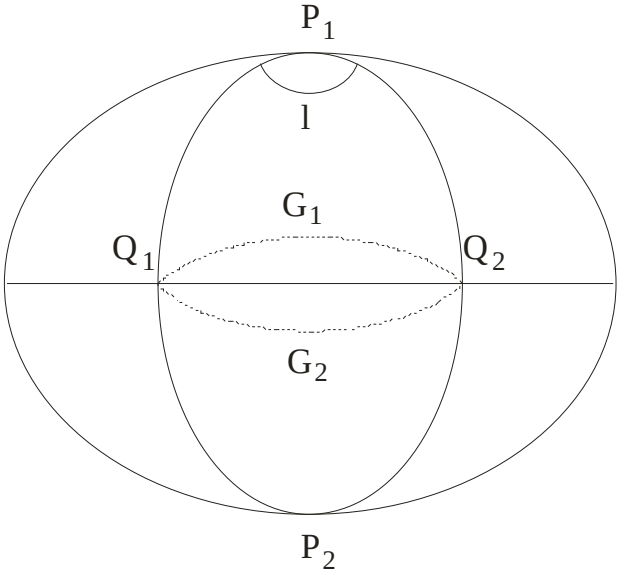


Рис. 7.36. Две геодезические линии

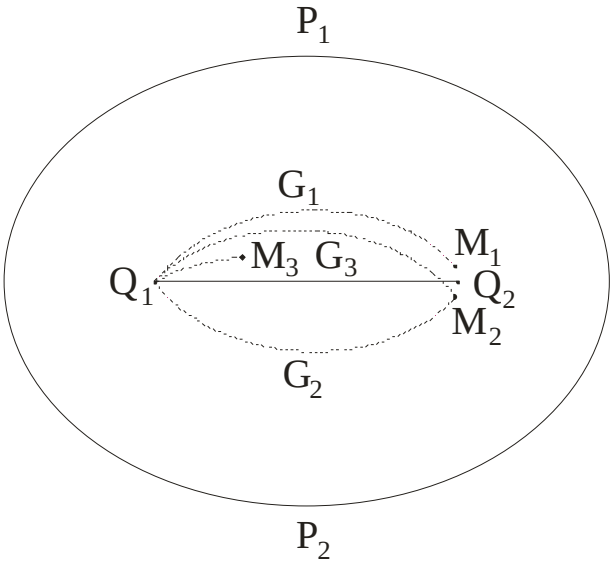


Рис. 7.37. Четыре геодезические линии

Более того. Дуга геодезической линии G_3 от точки Q_1 до точки ее пересечения с экватором (на рис. 7.37 она никак не обозначена) является геодезической линией между ними. Отрезок геодезической линии G_3 от точки пересечения с экватором до точки M_2 также представляет собой кратчайшую линию. Второй отрезок служит продолжением первого, но вместе они не образуют кратчайшую линию. При использовании центральных сечений эллипсоида подобные эффекты не имели бы места.

На рис. 7.36 можно заметить, что на экваторе в точках Q_1 и Q_2 прямое и обратное нормальные сечения совпадают, а геодезическая линия отклоняется от них на значительное расстояние. Если точка будет перемещаться по геодезической линии от Q_2 к Q_1, то в начале этого движения геодезическая линия также не будет проходить между взаимно обратными сечениями. Таким образом, положение геодезической линии относительно взаимно обратных нормальных сечений, представленное на рис. 7.31, имеет место только при определенных условиях.

Если бы мы измеряли в точке Q_1 угол между нормальными сечениями на точку M_1 и точку M_3, то мы могли бы наблюдать эффект, который можно назвать «эффектом геодезических линий»: правую точку M_3 мы видели бы слева, а левую точку M_1 – справа. Именно таким образом мы видим расположение точек при рассматривании рис. 7.37. Если рис. 7.37 дополнить центральными сечениями и геодезическими линиями, соединяющими вершины сфероидического треугольника $Q_1M_1M_3$, то получим картину, представленную на рис. 7.38, где геодезические линии обозначены штриховыми линиями. Нетрудно обнаружить, что обход вершин треугольника в порядке Q_1, M_1 и M_3 по геодезическим линиям будет происходить по часовой стрелке. Обход вершин в том же порядке по дугам центральных сечений будет осуществляться против часовой стрелки. Таким образом, если на эллипсоиде имеются два треугольника с одними и теми же вершинами, но стороны одного треугольника являются геодезическими линиями, а стороны другого – дугами центральных сечений, то возможны случаи, когда эти треугольники будут иметь противоположную ориентацию.

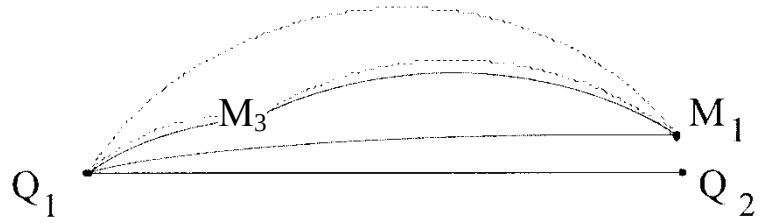


Рис. 7.38. Геодезические линии и центральные сечения

Центральные сечения эллипсоида обладают тем свойством, что через две диаметрально противоположные точки эллипсоида можно провести бесконечное множество таких сечений. Этим же свойством обладают большие круги на сфере, тогда как между двумя диаметрально противоположными точками эллипсоида существует только две (равные по длине) геодезические линии, каждая из которых проходит через один из полюсов. Единственным исключением из этого правила являются геодезические линии, соединяющие полюсы эллипсоида, которых можно провести любое количество. Между двумя точками эллипсоида, не лежащими на одном диаметре, можно провести только одно центральное сечение.

Таким образом, можно признать, что в качественном отношении центральные сечения, по крайней мере, не хуже геодезических линий. Теперь можно провести количественное сравнение этих кривых, для чего воспользуемся рис. 7.39. На нем слева изображен элементарный треугольник 1 на сфере с радиусом $R=1$, его гипотенуза ds является элементарной дугой большого круга. Справа на рис. 7.39 показаны треугольники 2 и 3 на эллипсоиде вращения, полученного линейным отображением сферы, следовательно, его малая полуось $b=R$. Гипотенуза dS_Γ треугольника 2 является элементарной дугой геодезической линии, а гипотенуза dS треугольника 3 – элементарной дугой центрального сечения эллипсоида вращения. Азимут геодезической линии на эллипсоиде и азимут большого круга на сфере равны и обозначены буквой A, азимут дуги центрального сечения эллипсоида обозначен греческой буквой α.

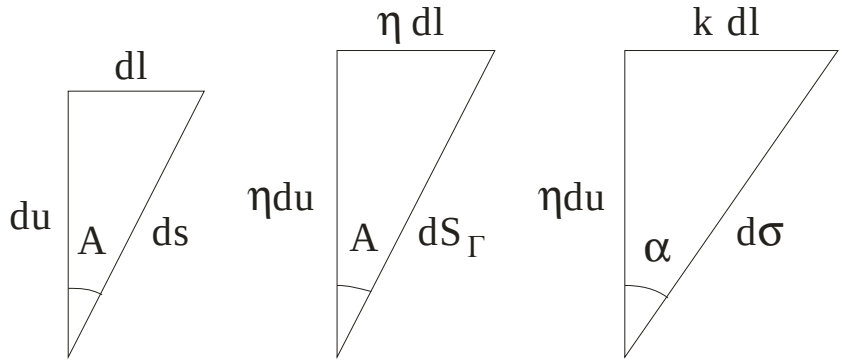


Рис.7.39. Элементарные треугольники

На эллипсоиде масштаб в меридианном направлении равен η, поэтому элементарный отрезок меридиана на эллипсоиде равен $\eta\, ds$. Тогда мы можем записать следующие равенства:

$$ds = \frac{du}{\cos A}; \qquad (7.228)$$

$$dS_\Gamma = \frac{\eta\, du}{\cos A}; \qquad (7.229)$$

$$dS = \frac{\eta\, du}{\cos \alpha}. \qquad (7.230)$$

На основании соотношений (7.214) и (7.228) последние два равенства можно представить как

$$dS_\Gamma = \eta\, ds; \qquad (7.231)$$

$$dS = m\, ds. \qquad (7.232)$$

Кроме того, соотношение (7.232) следует непосредственно из определения масштаба по направлению (7.205).

Значение масштаба

$$\eta = \sqrt{1 + t^2 \sin^2 u}$$

где $t = \mathrm{tg}\varepsilon$, с учетом формулы 8 из табл. 7.3 примет вид

$$\eta = \sqrt{1 + t^2 \cos^2 A_0 \sin^2 s}. \qquad (7.233)$$

Значение масштаба m по азимуту A представим следующим образом:

$$m^2 = \eta^2 \cos^2 A + k^2 \sin^2 A = \eta^2 + (k^2 - \eta^2) \sin^2 A$$

или

$$m = \sqrt{1 + t^2 \sin^2 A_0 + t^2 \cos^2 \mathrm{A}_0 \sin^2 s}, \qquad (7.234)$$

а также

$$m = k\sqrt{1 - e^2 \cos^2 \mathrm{A}_0 \cos^2 s}. \qquad (7.235)$$

Теперь мы имеем возможность записать следующие дифференциальные уравнения:

$$dS_\Gamma = \sqrt{1 + t^2 \cos^2 \mathrm{A}_0 \sin^2 s}\, ds; \qquad (7.236)$$

$$dS = \sqrt{1 + t^2 \sin^2 A_0}\sqrt{1 + \frac{t^2 \cos^2 \mathrm{A}_0}{1 + t^2 \sin^2 A_0} \sin^2 s}\, ds. \qquad (7.237)$$

Элементарная дуга dS может быть представлена также как

$$dS = k\sqrt{1 - e^2 \cos^2 \mathrm{A}_0 \cos^2 s}\, ds. \qquad (7.238)$$

Чтобы установить приближенное значение разности длины дуги центрального сечения эллипсоида и длины геодезической линии, применим традиционное разложение в ряд

$$\sqrt{1 + x} = 1 + \frac{1}{2}x - \frac{1}{2 \cdot 4}x^2 + \frac{1 \cdot 3}{2 \cdot 4 \cdot 6}x^3 - \frac{1 \cdot 3 \cdot 5}{2 \cdot 4 \cdot 6 \cdot 8}x^4 + \ldots$$

и ограничимся двумя первыми членами этого разложения:

$$dS_\Gamma = (1 + \frac{1}{2}t^2 \cos^2 \mathrm{A}_0 \sin^2 s)\, ds;$$

$$dS = k(1 - \frac{1}{2}e^2 \cos^2 \mathrm{A}_0 \cos^2 s)\, ds.$$

Тогда значение разности элементарных дуг dS и dS_Γ будет

$$dS - dS_\Gamma = k(1-\frac{1}{2}e^2\cos^2 A_0\cos^2 s)\,ds - (1+\frac{1}{2}t^2\cos^2 A_0\sin^2 s)\,ds$$

или

$$dS - dS_\Gamma = (k-1-\frac{1}{2}ke^2\cos^2 A_0)ds - \frac{1}{2}ke^2\cos^2 A_0\,(k-1)\sin^2 s\,ds .$$

После интегрирования данного выражения находим

$$S - S_\Gamma = (k-1-\frac{1}{2}k\,e^2\cos^2 A_0)s - \frac{1}{2}k\,(k-1)\,e^2\cos^2 A_0\,\frac{1}{2}(s-\sin s\cos s)$$

или

$$S - S_\Gamma = (k-1-\frac{1}{2}Q(k+1))s + \frac{1}{2}Q\,(k-1)\sin s\cos s , \qquad (7.239)$$

где

$$Q = \frac{1}{2}ke^2\cos^2 A_0 . \qquad (7.240)$$

Чтобы не анализировать не очень точное и не слишком изящное выражение для разности длин, откажемся от разложения в ряд, возведем в квадрат выражения (7.236) и (7.238) и возьмем их разность:

$$dS^2 - dS_\Gamma^2 = (k^2(1-e^2\cos^2 A_0\cos^2 s) - (1+t^2\cos^2 A_0\sin^2 s))\;ds^2 .$$

Преобразовав данное выражение с учетом $ke = t$, находим

$$dS^2 - dS_\Gamma^2 = t^2\sin^2 A_0\,ds^2$$

или

$$(dS + dS_\Gamma)(dS - dS_\Gamma) = t^2\sin^2 A_0\,ds^2 . \qquad (7.241)$$

Поскольку существуют два соотношения

$2ds \approx dS + dS_\Gamma$;

$2ds < dS + dS_\Gamma$,

постольку (7.241) можно представить как

$$2(dS - dS_\Gamma) < t^2\sin^2 A_0\,ds .$$

Проинтегрировав последнее выражение, находим

$$S - S_\Gamma < \frac{1}{2}t^2\sin^2 A_0\,s . \qquad (7.242)$$

Наибольшая разность длин центрального сечения и геодезической линии будет при $A_0 = 90^\circ$. Значение

$$\frac{1}{2}t^2 = \frac{1}{296{,}8}$$

примерно равно величине сжатия эллипсоида Красовского. Таким образом, разность длин геоцентрического сечения и геодезической линии пропорциональна величине сжатия эллипсоида. Наименьшее значение этой

разности имеет место при $A_0 = 0^\circ$, что очевидно, поскольку при указанном значении начального азимута геодезическая линия и центральное сечение совпадают.

7.16. Выбор координатного пространства в системах геомоделирования

Отличительной чертой геоинформационных систем является необходимость выполнения операций над данными о пространственном положении объектов и решения различных геометрических задач на земной поверхности. Существенное влияние на результаты их решения в ГИС оказывает выбор координатного пространства, или иначе – геометрической модели земной поверхности.

В настоящее время основным способом получения цифровых топографических карт и геоинформационных моделей остается картометрический, при котором цифровые карты или цифровые модели создаются по имеющимся обычным картам. Поскольку в нашей стране топографические карты составляются в проекции Гаусса – Крюгера, то цифровые топографические данные о пространственном положении объектов представляются в этой же проекции. В результате система плоских координат топографических карт трансплантируется в среду геоинформационных систем. Такой механический перенос координатного пространства является следствием инертности мышления и может быть сравним с тем, как конструкторы первых автомобилей копировали внешний вид карет. Сегодня аналогичным образом в ГИС копируется внешний вид обычных карт, включая геометрические искажения, зависящие от используемой картографической проекции.

Топографические карты предназначены не только для рассматривания картографического изображения, но и для непосредственных измерений по ним с помощью простейших инструментов. Различные проекции были разработаны для того, чтобы значения геометрических величин, измеренных по карте, были достаточно близки к значениям этих же величин на земной поверхности, и не возникала необходимость выполнения сложных для ручного способа вычислений. Таким образом, точность геометрических данных в ГИС ограничена, как правило, точностью исходных картографических материалов, хотя потенциально может быть более высокой.

Известно, что в проекции Гаусса – Крюгера углы искажаются на сравнительно небольшие величины, но искажения длин сторон весьма значительны. Так, на краях 6-градусных зон их относительные ошибки достигают величины 1 : 1 250, что во многих случаях является очень большой погрешностью. Например, относительная ошибка теодолитных ходов, прокладываемых в качестве планового обоснования крупномасштабных съемок, не должна превышать 1 : 2 000 от длины хода. По этой причине, а также в связи с вызывающими сомнения требованиями секретности, крупномасштабные съемки обычно выполняются не в государственной системе координат, как это представляется целесообразным с позиций здравого смысла, а в многочисленных системах условных координат.

Проблема выбора «координатного пространства» для систем геомоделирования разбивается на две задачи:

1) выбор модели геопространства – конкретной математической поверхности или трехмерного пространства;

2) определение системы координат на выбранной модели.

При выборе геометрической модели главными являются три критерия:

1) модель должна быть единой для всей территории страны, а в идеале – и для всей земной поверхности;

2) модель должна обеспечивать необходимую точность решения задач на любые расстояния и площади в пределах моделируемого геопространства;

3) модель должна быть удобной.

Принцип единства модели (или ее единственности) следует из экономических соображений: использование единой модели представляется более эффективным решением, поскольку отпадает необходимость согласования различных моделей и перехода от одних моделей к другим. Второе требование является очевидным: модели, не обеспечивающие необходимую точность, не имеют права на существование. Наконец, условие удобства применяемой модели есть не что иное, как еще одно ограничение, накладываемое на модель по экономическим соображениям.

Теоретически возможны пять вариантов геометрических моделей геопространства:

1) плоскость;

2) сфера;

3) эллипсоид вращения (либо трехосный эллипсоид);

4) трехмерное евклидово пространство с системой прямоугольных координат;

5) некоторая поверхность, близкая к эллипсоиду вращения.

Недостатки первого варианта отмечались выше. Но если в качестве модели земной поверхности все-таки выбрать плоскость, то сразу встает вопрос и о выборе картографической проекции. Поиски «хороших» картографических проекций продолжаются до сих пор, но необходимо признать, что возможности математической картографии практически исчерпаны, и в лучшем случае будут найдены картографические проекции, незначительно повышающие точность отображения достаточно ограниченных территорий. Исследования и разработки картографических проекций для ГИС, выполненные в [8], даже для сравнительно небольшой территории Белоруссии, располагающейся всего лишь в трех 6-градусных зонах, позволили довести эту погрешность только до 1 : 3 000. Для получения такой точности были приложены серьезные усилия и использован сложный математический аппарат, и улучшение результата если и возможно, то только на несущественную величину.

Второй вариант – сфера – может служить удовлетворительной моделью лишь для небольших по площади территорий. Ее применение в качестве модели поверхности всей Земли рассматривалось в [7], где приведены размеры сферы при наложении тех или иных условий. Так, если потребовать, чтобы погрешности длины дуги меридиана от экватора до полюса и длины четверти

дуги экватора были одинаковыми, то радиус сферы следует выбрать равным 6 372,9 км. Тогда относительная ошибка определения длин составит величину 1 : 1 200. Во многих приложениях такая ошибка не может быть признана допустимой. Но достоинством сферы является удобство решения на ней многих практических задач с использованием формул сферической тригонометрии.

Эллипсоид вращения сегодня служит эталоном модели земной поверхности: решения любых задач сравниваются с результатами их решения на сфероиде. Разногласия по этому вопросу сводятся лишь к определению параметров (размеров и формы) эллипсоида и ориентировки его в теле Земли. Но при всей простоте уравнения эллипсоида решение многих задач на нем характеризуется сложностью. Одна из таких задач – вычисление длин линий. Можно сказать, что над всей сфероидической геодезией висит проклятие эллиптического интеграла. И известные неудобства связаны не столько со сложностью вычислений, систематическим применением разложений в ряды, сколько с невозможностью получить *замкнутую систему формул*. Отчасти это вызвано применением геодезических линий. Но, как было показано выше, для решения главных геодезических задач такую систему формул получить можно, если использовать не геодезические линии на сфероиде, а центральные сечения.

Известны и другие методы решения главных геодезических задач. В 1954 г. М.С. Молоденским было получено строгое решение прямой и обратной геодезических задач в замкнутой форме с применением хорд эллипсоида, изложенное в [4]. Однако и после этого итеративный способ Бесселя считается основным при решении прямой геодезической задачи на большие расстояния.

Четвертый вариант модели геопространства – это трехмерное евклидово пространство с системой прямоугольных координат, центр которой совпадает с центром земного эллипсоида. Данная система используется в космической геодезии и стала применяться в фотограмметрии с началом фотосъемок Земли из космоса. Однако ее применение в системах геомоделирования осложняется одним обстоятельством: человеку трудно представить положение точки на земной поверхности по значениям ее трех прямоугольных координат. Поэтому, если получателем данных является человек, то такие данные в картографической или в табличной форме должны представляться в системе координат, определенной на земной поверхности.

Последний вариант – отказ от эллипсоида вращения и его замена другой поверхностью – может показаться большинству геодезистов слишком радикальным. Причин их предполагаемых возражений может быть две. Одна из них может сводиться к тому, что Земля при остывании должна была получить форму эллипсоида вращения. Да, могла, но не получила. В другой части высшей геодезии – физической геодезии – поверхностью Земли считается геоид – уровенная поверхность, совпадающая с поверхностью Мирового океана в спокойном состоянии и мысленно продолженная под материками таким образом, что отвесные линии совпадают с нормалями к ней. М.С. Молоденским было доказано, что поверхность геоида не может быть определена без знания распределения масс внутри Земли, и предложено использовать поверхность квазигеоида. Поверхность квазигеоида отступает от эллипсоида более чем на

100 м. Из формул (7.129) и (7.134) следует, что можно получить модель земной поверхности, отступления которой от эллипсоида вращения будут составлять менее 4 мм. Это не та величина, о которой можно говорить.

Другие потенциальные возражения могут носить экономический характер: переход к новой модели потребует дополнительных средств на переиздание карт, каталогов координат геодезических пунктов и т. п. Такие возражения также будут лишены оснований: реальная точность координат геодезических пунктов в пределах всей территории Российской Федерации составляет, в лучшем случае, несколько метров, а точность взаимного положения соседних пунктов – несколько сантиметров. Переход к поверхности, альтернативной эллипсоиду, практически никак не скажется на точности координат астрономо-геодезической сети. Точность топографических карт и планов еще ниже точности геодезических сетей. Поэтому и второй аргумент также можно игнорировать.

Логика обработки данных в системах геомоделирования требует их связи с банками геодезических данных, так как пункты геодезических сетей служат основой для выполнения всех дальнейших топографических и картографических работ. В высшей геодезии накоплен наиболее значительный опыт решения задач на земной поверхности, поскольку в этом и состоит ее назначение. Поэтому есть смысл рассмотреть методы решения задач при обработке геодезических сетей.

Прежде всего, следует отметить, что традиционная методика обработки геодезических сетей характеризуется следующими известными особенностями.

1. Стороны геометрических построений на поверхности эллипсоида считаются геодезическими линиями, хотя в действительности это не так.

2. Сфероидические треугольники считаются сферическими. При сторонах триангуляции менее 200 км значение разности сферического и сфероидического углов выражается величиной менее 0.001″.

3. Сферические треугольники решаются как плоские с применением способа Лежандра либо способа аддитаментов. При использовании *способа Лежандра* сферический треугольник решается как плоский, если каждый из его углов уменьшить на 1/3 сферического избытка. Последний для небольших треугольников вычисляют по приближенной формуле

$$\varepsilon = \frac{ab\sin C}{2R^2},$$

где a и b – стороны треугольника; C – угол между ними.

При решении сферических треугольников по способу аддитаментов стороны треугольников выражаются в угловой мере (как отношение длины стороны к радиусу). Аддитаментами называют поправки в длины сторон треугольников, вычисляемые по приближенным формулам

$$\left.\begin{aligned} A_a &= \frac{a^3}{6R^2} \\ A_b &= \frac{b^3}{6R^2} \\ A_c &= \frac{c^3}{6R^2} \end{aligned}\right\},$$

где a, b и c – стороны треугольника; R – радиус сферы, на которой расположен треугольник.

4. Главная геодезическая задача на большие расстояния решается с использованием приближенных методов, лучшим из которых считается способ Бесселя.

5. Уравнивание заполняющей геодезической сети выполняется на плоскости в проекции Гаусса – Крюгера.

В этой схеме все подчинено одной главной цели – максимальному упрощению последующих (самых массовых) вычислений при ручной обработке заполняющих геодезических сетей. Однако скорость выполнения вычислений современными компьютерами превышает скорость вычислений человеком примерно на 10 порядков. Поэтому вычислительная сложность обработки геодезических сетей уже не является определяющим критерием. На первое по важности место среди свойств методов вычислений в настоящее время следует поставить точность вычислений. Кроме того, в условиях дефицита и высокой стоимости программных средств для обработки геодезических сетей важную роль играет универсальность алгоритмов. Однако в учебниках по сфероидической геодезии можно обнаружить множество приближенных формул для различных частных случаев: формулы для расстояний до 20 км, до 40 км, до 60 км, до 200 км и т. д.

Общим свойством описанной выше схемы обработки геодезических сетей служит то, что очень многие формулы являются приближенными и получены в результате разложения в ряды. Можно заметить, что интенсивное использование разложений в ряды является характерной чертой сфероидической геодезии. В больших геодезических сетях отбрасывание всех членов ряда с малыми значениями означает накопление систематических ошибок.

Как сообщалось в [5] и [10], трудно дать объяснения некоторым фактам, полученным в результате повторного уравнивания астрономо-геодезической сети (АГС). Так, поправки в углах заполняющей сети 2-го класса в некоторых случаях достигали 10–15″. Заметим, что углы в ней измерялись с очень высокой точностью: в соответствии с инструкцией в сети триангуляции 2-го класса невязки в треугольниках не должны были превышать 4″, а средняя квадратическая ошибка измерения углов, вычисленная по невязкам

треугольников, не должна была превышать 1″. Таким образом, поправки в углы иногда более чем на порядок превышают значение средней квадратической погрешности измерения углов. Кроме того, при переуравнивании АГС смежные блоки из многих тысяч пунктов получали разворот относительно друг друга на величины от –0,5 до +0,27″ [5, 10].

Необходимо сказать, что проблема уравнивания больших геодезических сетей – старая проблема. Автор данной книги был еще студентом, когда в Новосибирский институт инженеров геодезии, аэрофотосъемки и картографии в конце 1960-х гг. приезжал П.А. Гайдаев с лекцией, посвященной проблемам уравнивания АГС. В своей лекции он сообщил, что при первом уравнивании АГС поправки в углы триангуляции 2-го класса иногда достигали 8″, что никак не согласуется с точностью измерения углов. По ходу лекции П.А. Гайдаев сделал замечание, что если бы наблюдатели знали об искажениях измеряемых величин в результате уравнивания, то они не осуществляли бы измерения так тщательно, как предписывается инструкциями. По мнению П.А. Гайдаева, порок заключался в методологии уравнивания больших геодезических сетей: ряды триангуляции 1-го класса не могли служить основой для заполняющих сетей 2-го класса, поскольку точность сплошной сети 2-го класса оказывалась выше, чем точность рядов триангуляции 1-го класса. В свое время это было доказано К.Л. Проворовым в его кандидатской диссертации, за которую ему была присвоена ученая степень доктора технических наук.

К сказанному можно добавить соображения П.С. Закатова, приведенные им в [4, с. 394] в короткой главе, посвященной уравниванию АГС: «Но если строгим уравниванием нельзя сделать сеть, имеющую невысокую точность полевых измерений, более точной, то хорошую в полевом исполнении сеть можно испортить применением неправильных методов и приемов ее обработки. Поэтому, учитывая значение астрономо-геодезической сети, необходимо для ее математической обработки применять продуманную и научно обоснованную программу и методику. Недостатки математической обработки астрономо-геодезической сети проявятся при обработке сетей всех последующих классов. Эти недостатки, в виде дополнительных ошибок исходных данных, отрицательно повлияют на точность всех геодезических сетей последующих классов».

Заканчивал же П.С. Закатов эту главу совсем уж пессимистически: «Не исключен в дальнейшем при некоторых обстоятельствах и отказ от уравнивания вообще; при соответствующей точности полевых данных уравнивание может оказаться не только ненужным, но даже вносящим некоторые дополнительные искажения или погрешности в результативные данные» [4, с. 398].

Как видим, наихудшие предположения П.С. Закатова сбылись, и за прошедшие 30–40 лет поправки, так беспокоившие П.А. Гайдаева, увеличились примерно в 1,5 раза.

Нельзя утверждать категорически, но можно высказать осторожное предположение, что указанные и некоторые другие эффекты являются следствием не только несовершенства методов уравнивания, но также способа

передачи координат по сети, в том числе и отмеченного накопления систематических ошибок.

Применение рядов обусловлено трактовкой сторон треугольников как геодезических линий. Использование сторон объяснялось соображениями удобства: «Если решать треугольники по обычным формулам сферической тригонометрии, то стороны необходимо выражать в частях радиуса, но это неудобно, так как практически стороны должны быть выражены в метрах» [4, с. 68].

Таким образом, возможность использования центральных углов, стягивающих стороны геодезических построений, теоретически допускалась, но отвергалась по практическим соображениям. Сегодня эта аргументация не представляется убедительной. Стороны сферических треугольников вполне можно выражать не в линейной, а в угловой мере. И неудобство здесь заключается только в том, что мы имели бы дело с очень малыми углами (до нескольких десятков угловых минут). Но даже при ручной обработке их можно было бы выражать в долях секунд, а после производства уравнительных вычислений переводить стороны из угловой меры в линейную.

При выполнении вычислений на компьютере пользователь видит только входные и выходные данные. Внутреннее представление данных от него скрыто и поэтому может осуществляться в любых удобных для вычислительной машины мерах и единицах измерений. Появление электронных вычислительных машин сразу повлияло на геодезические вычисления: до этого разные авторы любили подчеркивать, что полученные ими формулы удобны для приведения к логарифмическому виду. В период 1960-х гг. необходимость в таких формулах отпала.

Другой причиной необходимости переосмысления методов решения задач сфероидической геодезии являются изменения, произошедшие в методах геодезических построений, а именно, массовый переход к электронным методам измерения расстояний и, что более важно, к использованию методов автономного определения координат. В [10] совершенно справедливо отмечалось, что в геодезии наступает новая эпоха.

В сущности, такая эпоха уже наступила. Использование спутниковых систем определения координат стало массовым. Но если в высшей геодезии принято представлять положение точек в системе геодезических координат, а геоцентрическая или приведенная широта используются, как правило, только в процессе вывода тех или иных формул, то в космической геодезии используется геоцентрическая система координат. Результаты измерений астрономо-геодезической сети будут в течение продолжительного времени обрабатываться совместно с результатами спутниковых определений координат, поэтому следует считать целесообразным применение для этой цели единой системы координат. При определении единого координатного пространства систему геоцентрических координат следует предпочесть как более перспективную. Вообще же связь между приведенной, геоцентрической и геодезической широтами является простой и поэтому проблема выбора системы координат не

является столь принципиальной, как выбор поверхности или пространства в качестве модели.

Кроме того, переход к методам спутниковой геодезии позволяет существенно повысить точность определения взаимного положения точек на земной поверхности. Этот факт, в свою очередь, требует существенного повышения точности вычислений.

Наконец, следует отметить, что еще одним стимулом к совершенствованию методов решения задач на эллипсоиде служат потребности динамической геодезии. Последняя становится реальностью вследствие тех преимуществ, которые появляются в связи с внедрением методов космической геодезии: повышение точности измерений и производительности труда. Необходимость определения малых перемещений земной поверхности также требует повышения точности вычислений. Использование методов обработки геодезических измерений, в результате которых поправки в измеренные величины на порядок превышают погрешности их измерений, не может обеспечить возможность надежной фиксации достаточно тонких эффектов при изучении движений земной коры на значительных расстояниях и/или малых интервалах времени.

Для систем геомоделирования наиболее естественным координатным пространством представляется система координат, связанная с земной поверхностью, а не с картографической проекцией. В этом случае любая такая система позволяет легко интегрировать данные об объектах, расположенных на сколь угодно большом удалении. Поэтому, по мнению автора, унификация и стандартизация цифровых топографических данных должна допускать их представление только в системе координат, связанной непосредственно с земной поверхностью.

Перечисленный выше ряд проблем может быть решен, если:

1) в системах геомоделирования отказаться от хранения пространственных данных в системах координат картографических проекций и решение геометрических задач осуществлять непосредственно на поверхности земного эллипсоида либо другой, близкой к нему поверхности;

2) в качестве системы координат использовать систему геоцентрических координат;

3) вместо геодезических линий использовать центральные сечения;

4) по измеренным значениям сторон геодезических построений вычислять стягивающие их центральные углы и решение задач осуществлять по предложенным в настоящей главе формулам либо другим.

Отказ от плоскости и переход к решению задач на поверхности эллипсоида – это качественно новый этап в развитии геоинформационных систем, внедрение в ГИС методов сфероидической геодезии.

Сфероидическая геодезия представляется вполне сформировавшейся теоретической дисциплиной, созданной усилиями и трудом многих исследователей. Но если задать вопрос о необходимости и возможности ее дальнейшего развития, то на него можно ответить положительно. Основанием

для подобного утверждения служат не только перечисленные выше внешние, но и внутренние причины.

К внутренним причинам необходимости развития сфероидической геодезии можно отнести ее избыточную сложность и недостаточную последовательность. Сфероидическая геодезия является сильно математизированной или даже математической теорией, приложением геометрии к решению задач на эллипсоиде. Отсюда вытекает необходимость учитывать изменения, происходящие в математике.

Французским математиком Ж. Дьедонне были высказаны очень важные соображения, заслуживающие того, чтобы быть процитированными полностью: «Математики предыдущих столетий, а в еще большей мере философы и очеркисты, писавшие о математике, слишком хорошо преуспели в том, чтобы внедрить в сознание ″культурного″ общества образ застывшей и неизменной науки, восседающей на Олимпе ″абсолютных истин″ и благочестиво передаваемой от поколения к поколению без изменения в ней хотя бы одной буквы. Эта наука якобы не знает мук поиска и неясностей бедных наук, называемых ″экспериментальными″. Уже более ста лет тому назад математики-профессионалы освободились от столь наивных амбиций, но, видимо, понадобятся многие годы упорного труда, чтобы повсеместно прикончить подобные штампы: нет ничего более трудного, чем выкорчевывание ″внедрившихся идей″.

Источник недоразумения состоит в следующем. Действительно, теоремы, доказанные 2000 лет тому назад, верны сегодня в той же мере, как и в момент их открытия, между тем как ″экспериментальные истины″ возникают всегда лишь как некоторые приближения к действительности, нуждающиеся в постоянном усовершенствовании. Но то, что в математике, как и во всякой науке, меняется – это точка зрения, с которой рассматриваются ранее полученные результаты, причем, как и во всех науках, такие изменения происходят в настоящее время с возрастающей скоростью. Если отвлечься от несущественных признаков, то развитие математики ничем не отличается от развития других наук: новые открытия и их обсуждения приводят к необходимости переосмысливания старых теорем, к необходимости исследовать их взаимное отношение в свете новых теорий с тем, чтобы вписать их самым рациональным образом в новый контекст. Собственно для математики при этом характерно, что при подобных периодических переворотах сами теоремы остаются в целостности, а не распадаются на более точные. Они не могут быть опровергнуты более совершенной экспериментальной техникой, как это происходит с самыми твердо установленными ″фактами″ физики и биологии. Зато с величественного пьедестала ″основных теорем″ они зачастую спускаются в подчиненное положение ″следствий″, все менее и менее значительных, чтобы закончить свое существование в кладовке ″упражнений″, оставляемых для тренировки учащегося. Осознание этого установившегося исторического процесса развития должно привести математика-профессионала к более скромному пониманию своей роли: возможно, что открытия, которые ему иногда стоили стольких усилий и которыми он, вполне естественно, хотел

бы гордиться, рискуют превратиться в дальнейшем в игрушки для школьников будущих поколений» [2, с. 10–11].

К словам Дьедонне следует добавить, что по мере накопления новых данных, знаний и требований практики требуется переосмысление и прикладных теорий, в том числе – сфероидической геодезии. Одним из таких важных процессов, происходящих в настоящее время внутри математики и представляющих интерес в связи с обсуждаемым вопросом, является ее линеаризация, проникновение идей и методов линейной алгебры в различные области математики.

В данной главе показано, что применение линейной алгебры при решении некоторых задач сфероидической геодезии позволяет:

– сделать теорию более простой и более последовательной;

– дать простую и ясную интерпретацию таких параметров, как эксцентриситет эллипса, первая и вторая геодезические величины;

– получить с применением геоцентрических сечений эффективное решение прямой и обратной геодезических задач на эллипсоиде в виде замкнутых формул.

Еще одним аргументом в пользу применения методов линейной алгебры в сфероидической геодезии служит противопоставление «геометрии в целом» и «геометрии в малом». Предметом изучения геометрии в целом, как следует из ее названия, являются геометрические объекты (кривые, поверхности, многообразия) на всем их протяжении. Геометрия в малом занимается изучением свойств сколь угодно малых частей геометрических объектов. Такое разделение геометрии возникло в немецкой математической литературе в начале XX в., когда было обнаружено, что методы классической дифференциальной геометрии, в силу их локального характера, недостаточны для изучения геометрических объектов в целом. Выше было показано, что с точки зрения линейной алгебры эллипсоид представляет собой определенное отображение сферы. Можно сделать вывод, что дифференциальные свойства эллипсоида в известной степени объясняются свойствами линейного преобразования трехмерного пространства в целом, при котором сфера трансформируется в эллипсоид. Поэтому линейную алгебру можно рассматривать как одно из направлений геометрии в целом.

Но линейная алгебра никоим образом не может полностью исключить дифференциальную геометрию, поэтому следует говорить об их оптимальном сочетании при изложении теории сфероидической геодезии, или, выражаясь словами Дьедонне «вписать их самым рациональным образом в новый контекст».

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Бессель Ф.В. Избранные геодезические сочинения. – М.: Изд. геодезической лит., 1961. – 282 с.
2. Дьедонне Ж. Линейная алгебра и элементарная геометрия. – М.: Наука, Глав. ред. физ.-мат. лит. 1972. – 336 с.
3. Градштейн И.С., Рыжик И.М. Таблицы интегралов сумм, рядов и произведений. – М.: Наука, Глав. ред. физ.-мат. лит., 1971. – 1108 с.
4. Закатов П.С. Курс высшей геодезии. – М.: Недра, 1975. – 511 с.
5. Кашин Л.А. Построение классической астрономо-геодезической сети России и СССР (1816–1991 гг.). – М.: Картгеоцентр – Геодезиздат, 2001. – 192 с.
6. Математический энциклопедический словарь. – М.: Сов. энциклопедия, 1988. – 847 с.
7. Морозов В.П. Курс сфероидической геодезии. – М.: Недра, 1979. – 296 с.
8. Подшивалов В.П. Теоретические основы формирования координатной среды для геоинформационных систем. – Новополоцк: Изд. ПГУ, 1998. – 125 с.
9. Справочник геодезиста (в двух книгах). – М.: Недра, 1975. – 1056 с.
10. Хаимов З.С. К выходу в свет книги Л.А. Кашина // Геодезия и картография. – 2000. – № 5. – С. 48–52.
11. Кравченко Ю.А. Об интерпретации обозначении параметров в сфероидической геодезии // Геодезия и картография. – 2000. – № 4. – С. 25–28.
12. Кравченко Ю.А. Вычисление длины дуги меридиана // Геодезия и картография. – 2000. – № 5. – С. 8–12.
13. Кравченко Ю.А. Интерпретация параметров эллипса с позиций проективной геометрии // Геодезия и картография. – 2000. – № 10. – С. 18–25.
14. Кравченко Ю.А. Решение главной геодезической задачи на эллипсоиде // Геодезия и картография. – 2000. – № 2. – С. 45–51.

8. МОДЕЛИРОВАНИЕ ТОПОГРАФИЧЕСКИХ ПОВЕРХНОСТЕЙ

Обязательным компонентом функционально полных автоматизированных систем картографирования (АСК), автоматизированных картографических систем (АКС), геоинформационных систем (ГИС) или систем автоматизированного проектирования (САПР) объектов строительства является подсистема моделирования высот земной поверхности, качество функционирования которой в значительной мере влияет не только на стоимость информационных моделей местности (ИММ) в целом и получаемых на их основе карт или планов, но также на качество проектных и иных решений.

Для снижения расходов на создание геоинформационных моделей необходимо более эффективно использовать такой сложный и дорогостоящий инструмент, как вычислительный комплекс. С инженерной точки зрения, исследование методов информационного моделирования некоторого фрагмента реального мира означает выбор оптимального способа конкретного применения ЭВМ. Вообще же, чтобы получить правильное представление об эффективности геоинформационного моделирования, его оценку необходимо производить с учетом тех преимуществ, которые получают потребители, используя геоинформационные модели.

Таким образом, развитие топографо-геодезического и картографического производства вызвано как внутренними причинами – необходимостью непрерывного повышения его эффективности, так и внешними – необходимостью более полного удовлетворения информационных потребностей пользователей. Пока что экономический эффект получают преимущественно потребители, имеющие возможность принимать оптимальные решения на основе применения геоинформационных моделей и соответствующего программного обеспечения.

В области информационного моделирования рельефа практически отсутствует достаточно разработанная и общепризнанная система понятий. Сюда входит множество дисциплин, число и тип которых зависят от окончательного применения информационных моделей. Некоторые из этих дисциплин имеют очень мало общего с местностью. Хотя некоторые термины определены ГОСТ, положение не изменилось до сих пор, и почти каждый пишущий по этой проблеме пользуется собственной терминологией.

Таким образом, первая проблема в сфере информационного моделирования земной поверхности – это терминология. Проблема точности научных понятий рассматривалась в коллективной монографии [30]. Краткие выводы сводятся к следующему. При разработке сложных систем, описании их структуры и функционирования часто невозможно ограничиться языком, точность которого соответствует точности современного физико-математического знания. В связи с этим различают точные и размытые (диффузные) понятия. Чем менее строго определено понятие, тем большее значение имеет семантика названия, так как при интерпретации такого понятия неизбежно обращение к семантическим ассоциациям. Новые термины должны отвечать требованию семантической корректности, что означает отсутствие использования этого или родственного

понятия в сходных ситуациях в другом смысле. Диффузные понятия должны соответствовать принципу локальной проверяемости. Понятие считается противоречащим принципу локальной проверяемости, если может быть указана или хотя бы придумана ситуация, в которой понятие противоречиво или не поддается проверке.

Реальные объекты, процессы или явления обладают очень большим числом свойств и отношений. При моделировании некоторой предметной области необходимо четко представлять, какие стороны действительности воспроизводятся и какова адекватность модели. Поэтому определение основных понятий уместно начинать с предмета моделирования.

8.1. Топографическая поверхность

Когда говорят о моделировании рельефа с помощью ЭВМ, то часто имеют в виду лишь один его аспект – положение точек земной поверхности в пространстве. Какие-либо другие качества рельефа в его информационных моделях обычно не воспроизводятся, поскольку значения других геометрических характеристик рельефа могут быть получены как их функции. Исключения из этого правила довольно редки, ими являются, например, модели крутизны склонов. Но термины «рельеф» и «земная поверхность» используются неоднозначно.

В геоморфологическом понимании рельеф представляет собой совокупность форм вертикального и горизонтального расчленения земной поверхности. Таким образом, рельеф при этом трактуется как некоторое многообразие элементарных первичных форм, в совокупности образующих (вторичную)* земную поверхность в целом. Однако при изучении ее геометрической формы в целом оказывается результативным такой подход, когда земная поверхность считается некоторой математической «поверхностью относимости» (сферой, эллипсоидом вращения или трехосным эллипсоидом), на которой определена та или иная система координат, не обязательно прямоугольных. Такова геодезическая трактовка земной поверхности.

Точки математической поверхности отличаются своими координатами, поэтому можно считать, что координаты (x, y) идентифицируют точки как элементы бесконечного множества – поверхности относимости. Каждой точке математической поверхности присуще определенное качество, свойство, атрибут – высота, расстояние от этой точки до реальной физической поверхности. Таково топографическое понимание земной поверхности.

Кроме того, в зависимости от конкретной задачи под поверхностью Земли, как физического тела, понимают либо поверхность суши и дна океанов, либо только поверхность суши.

Вместе с тем, понятие рельефа не эквивалентно понятию земной поверхности, если его рассматривать как элемент содержания топографических карт и планов. На топографических картах и планах рельеф изображается с

---

* В данном случае и далее по тексту в подобных конструкциях скобки ставятся автором с целью пояснения информации, которая может быть непонятна читателю. – Прим. ред.

помощью условных знаков, горизонталей и подписей значений высот отдельных точек земной поверхности. При этом к рельефу относят такие элементы местности, которые не являются точками или участками земной поверхности (выходы подземных газов, отдельные камни или скопления камней, служащие ориентирами и т. п.). Изображение таких элементов на картах не требует воспроизведения поверхности, и задача построения их картографических изображений принципиально не отличается от задачи построения изображений элементов ситуации.

Хотя основным средством изображения поверхности Земли как физического тела на топографических картах служат горизонтали, некоторые участки земной поверхности с их помощью при заданном масштабе и/или сечении иногда не могут быть изображены, либо их не принято изображать. К таким участкам относят овраги, обрывы, промоины и т. д. Кроме того, при изображении рельефа на топографических картах и планах при помощи цвета воспроизводится еще одно его свойство – происхождение. Элементы рельефа естественного происхождения показываются коричневым цветом, искусственного – черным и только при помощи условных знаков (дамбы, насыпи и т. д.) и подписей высот. Участки естественного рельефа с нарушениями антропогенного характера при помощи горизонталей не отображаются, а оконтуриваются и сопровождаются пояснительной подписью.

Наконец, можно указать на известные различия в интерпретации рельефа картографами и некоторыми категориями потребителей топографической и картографической продукции, например, проектировщиками. Картографы, как и геоморфологи, склонны понимать под рельефом, в первую очередь, естественный рельеф. Проектировщиков же история формирования земной поверхности интересует в меньшей степени, так как при решении их задач (допустим, при определении пространственного положения проектируемого сооружения) этот фактор часто не имеет значения.

Учитывая все вышесказанное, в дальнейшем вместо термина «рельеф» будем использовать понятие топографической поверхности, имея в виду реально существующую или проектируемую физическую земную поверхность. Использование этого термина преследует две цели:

– подчеркнуть, что речь идет только о геометрической форме физической земной поверхности, и другие аспекты рельефа не рассматриваются;

– указать на топографическую интерпретацию понятия земной поверхности.

Ранее этот термин использовался в работах [11, 17, 18, 31, 33].

8.2. Структурные линии и точки

Участок топографической поверхности достаточно больших размеров содержит такие элементы, как *структурные линии* и *структурные точки*. Хотя эти понятия используются часто, они довольно расплывчаты. Одним из авторов структурная линия квалифицировалась, например, как «линия количественного или качественного изменения закона интерполяции высот, совпадающая с реально существующими на местности границами между отдельными

элементами рельефа». Но любая линия на поверхности связана с изменением чего-нибудь, следовательно, под это определение, в сущности, попадают все линии на любой поверхности, не являющейся поверхностью постоянной кривизны. И таких линий даже на малых участках подобных поверхностей – бесконечное множество. Очевидно, что данное определение весьма расплывчато.

Иногда структурные линии определяются перечислением, например, в [8] к ним отнесены линии водоразделов, тальвегов и подошв. В Справочнике геодезиста [28] *характерными*, то есть структурными линиями, названы водоразделы и тальвеги, а характерными точками – точки вершин, дна котловин и седловые точки. В дополнение к перечисленным характерным точкам А.С. Васмут различал еще угловые, устьевые, развилочно-тальвеговые, мысовые, поворотные, точки урезов воды (?) и точки пересечения: подошвы с тальвегами, подошвы с водоразделами и т. д. Комментировать эти определения не будем.

Вопрос заключается в том, каким образом определить то общее свойство, которое позволяет выделить из бесконечного множества линий и точек топографической поверхности класс структурных линий и структурных точек. Ниже это свойство определяется с помощью некоторых понятий дифференциальной геометрии. Последняя занимается изучением внутренней геометрии поверхностей, представляющей совокупность свойств, не зависящих от выбора системы координат. Поскольку внутренняя геометрия поверхности никаким образом не изменяется при ее перемещениях и поворотах в пространстве, постольку привлечение понятий дифференциальной геометрии представляется разумным.

Соприкасающейся окружностью C плоской кривой *L* в ее точке *M* называется предельное положение окружности, проходящей через *M* и две ее соседние точки M_1 и M_2 при их стремлении к *M* (рис. 8.1). Центр *O* этой окружности называют *центром кривизны* кривой *L* в точке *M*. Радиус *r* соприкасающейся окружности называют *радиусом кривизны* кривой *L* в точке *M*, а величина, обратная радиусу кривизны

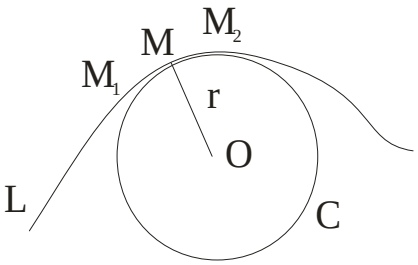


Рис. 8.1. Соприкасающаяся окружность

$$k = \frac{1}{r}, \qquad (8.1)$$

определяется как *кривизна* кривой *L* в точке *M*.

Формально *кривизна плоской кривой* в точке *M* определяется как предел отношения

$$k = \lim_{\Delta s \to 0} \frac{\Delta \alpha}{\Delta s},$$

где $\Delta\alpha$ – приращение угла касательной; Δs – приращение длины дуги кривой. Для вычисления кривизны плоской кривой используется формула

$$k = \frac{f''(x)}{\left(1 + (f'(x))^2\right)^{3/2}}. \qquad (8.2)$$

Прежде чем рассматривать структурные линии на топографической поверхности, рассмотрим характерные точки на кривых. Если попросить кого-либо выбрать точки на кривой (рис. 8.2) для последующей ее отрисовки и больше задание никак не уточнять, то вероятнее всего, что большинство выберет точки показанные более жирно. Во всяком случае, именно так поступят топографы, имеющие опыт съемки. Если затем попросить уточнить положение кривой и указать дополнительные точки, то будут выбраны точки, отмеченные не так жирно.

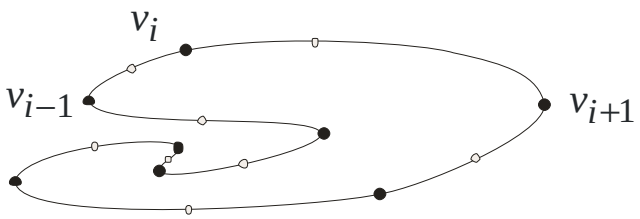


Рис. 8.2. Характерные точки кривой:

○ точки минимальной кривизны;

• точки максимальной кривизны

Общим свойством выбранных точек первого типа является то, что в окрестности каждой из них кривизна плоской кривой достигает своего максимального значения. В окрестности каждой точки второго типа кривизна кривой имеет минимальное значение. Таким образом, в выбранных точках кривизна принимает экстремальные значения.

Точки максимальной и минимальной кривизны на гладкой кривой будут чередоваться. Если на кривой выбрать три последовательные точки v_{i-1}, v_i, v_{i+1} с экстремальным значением кривизны, то максимальное отклонение кривой от хорды, стягивающей крайние точки v_{i-1} и v_{i+1}, будет находиться вблизи средней точки v_i.

Множество точек P(x, y, z), координаты которых удовлетворяют уравнениям

$$\left.\begin{aligned} x &= x(u,v) \\ y &= y(u,v) \\ z &= z(u,v) \end{aligned}\right\}, \qquad (8.3)$$

где u и v – действительные числа, называется непрерывной поверхностью, если x, y и z являются непрерывными функциями параметров u и v. Поверхность может быть также определена уравнением

$$f(x, y, z,) = 0 \qquad (8.4)$$

или

$$z = z(x, y). \qquad (8.5)$$

Точка P поверхности называется регулярной точкой, если при параметрическом задании поверхности в достаточно малой окрестности P

функции (8.3) имеют непрерывные частные производные первого порядка и хотя бы один из определителей

$$\begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} \\ \frac{\partial x}{\partial v} & \frac{\partial y}{\partial v} \end{vmatrix}, \quad \begin{vmatrix} \frac{\partial y}{\partial u} & \frac{\partial z}{\partial u} \\ \frac{\partial y}{\partial v} & \frac{\partial z}{\partial v} \end{vmatrix}, \quad \begin{vmatrix} \frac{\partial z}{\partial u} & \frac{\partial x}{\partial u} \\ \frac{\partial z}{\partial v} & \frac{\partial x}{\partial v} \end{vmatrix}$$

отличен от нуля.

Нормалью к поверхности S в ее регулярной точке P называется прямая, проходящая через P и перпендикулярная к касательной плоскости в этой точке. Нормальным сечением называют сечение поверхности плоскостью, содержащей нормаль к поверхности. Все множество точек поверхности принято делить на омбилические и неомбилические. Если в некоторой точке существуют два нормальных сечения, которым соответствуют наибольшая $k_{\max}$ и наименьшая $k_{\min}$ величины кривизны, то такие сечения называются главными нормальными сечениями, а такая точка – неомбилической. Плоскости главных нормальных сечений взаимно перпендикулярны. Точку, в которой кривизны всех нормальных сечений k_n равны между собой, называют омбилической. Существуют поверхности постоянной кривизны, то есть поверхности, все точки которых являются омбилическими (плоскость и сфера).

Линией кривизны на поверхности называется кривая, в каждой точке которой касательная принадлежит плоскости главного нормального сечения в этой точке. Таким образом, через каждую неомбилическую точку поверхности проходят две линии кривизны, которые всегда являются перпендикулярными. И линии кривизны покрывают всю поверхность, за исключением омбилических точек. Изолированные омбилические точки можно рассматривать как вырожденные, стянутые в точку замкнутые линии кривизны. В этом смысле омбилические точки похожи на точки локальных экстремумов на поверхности, которые можно трактовать как вырожденные замкнутые изолинии.

Очевидно, что структурная линия является линией кривизны, т. е. такой кривой, в каждой точке которой касательная принадлежит плоскости главного нормального сечения в этой точке. Поскольку через любую точку выделенной на поверхности линии кривизны по нормали к последней проходит другая линия кривизны, которая может и не быть структурной линией, то очевидно, что определение структурных линий, как линий кривизны, нуждается в уточнении.

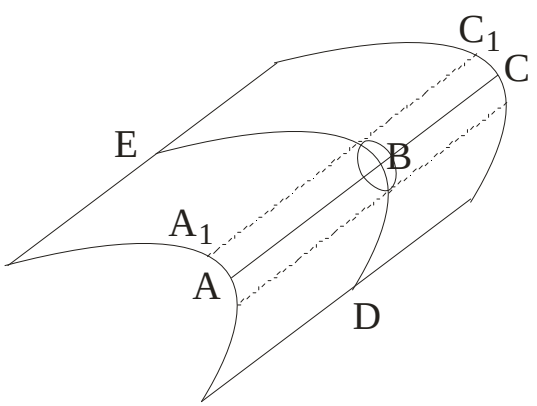


Рис. 8.3. Структурная линия

С этой целью рассмотрим участок поверхности, на котором имеется структурная линия AC (рис. 8.3). Пусть B

– произвольная точка структурной линии АС. Выделим сколь угодно малую окрестность точки В. Главное нормальное сечение, проходящее через касательную к структурной линии АС в точке В, обозначим N_p, а главное нормальное сечение в той же точке, проходящее через касательную к линии кривизны DE, – через N_q. Обозначим кривизны главных нормальных сечений соответственно через k_p и k_q.

Точка В отличается от любой другой точки своей окрестности, не принадлежащей структурной линии АС, тем, что кривизна k_q главного нормального сечения N_q достигает в ней экстремального значения. Если в полосе, сколь угодно близкой к структурной линии, провести некоторую линию кривизны A_1C_1, не совпадающую со структурной линией АС, то ее точки таким свойством обладать не будут. Далее структурной линией будем называть линию кривизны, в каждой точке которой кривизна k_q (скорость вращения касательной плоскости вокруг кривой) имеет экстремальное значение по сравнению с k_q в точках другой линии кривизны, сколь угодно близкой к первой.

Данное определение позволяет иначе взглянуть на состав класса «структурные линии». По каким-то причинам структурными называют, как правило, только определенные гладкие линии и только на естественном рельефе. Кроме того, обычно негласно полагается, что кривизна в точках структурной линии максимальна, как это имеет место на рис. 8.3.

Рассмотрим участок U-образной долины, изображенный на рис. 8.4, и его профиль по линии АВ. В данном примере структурными линиями являются линии подошв склонов и линия тальвега, хотя последняя и не является совокупностью точек с максимальной кривизной k_q.

Положение несколько усложняется, если рассматривать противоположный пример (рис. 8.5). Таким «языком» может заканчиваться более или менее вытянутый хребет. Линия водораздела АОВ попадает в разряд структурных линий поверхности в традиционном толковании, но опять-таки не все ее точки обладают свойством максимальной кривизны k_q. Две другие структурные линии ОС и OD не подходят под определение структурной линии в традиционном понимании, так как не являются ни линиями подошвы, ни совокупностью точек «изменения крутизны склонов» (рис. 8.6).

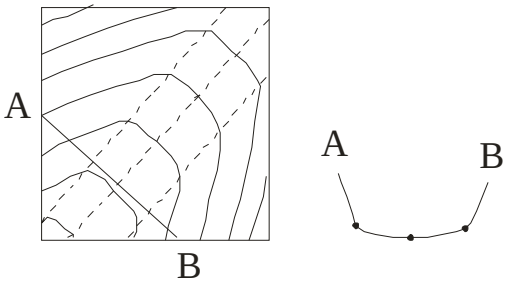


Рис. 8.4. U-образная долина

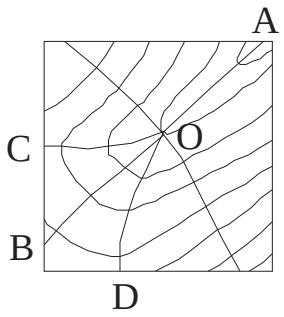


Рис. 8.5. Водораздел

Обе указанные линии OC и OD можно было бы назвать водоразделами второго порядка, но тогда возникает вопрос: чем отличаются водоразделы первого и второго порядков? Чтобы выяснить это, обратимся к рис. 8.7. Если кривые OA, OB, OC назвать водоразделами первого порядка, а OE, OF, OG – водоразделами второго порядка, то первые являются такими структурными линиями, у которых кривизна k_q главного нормального сечения N_q максимальна, а у водоразделов второго порядка она минимальна.

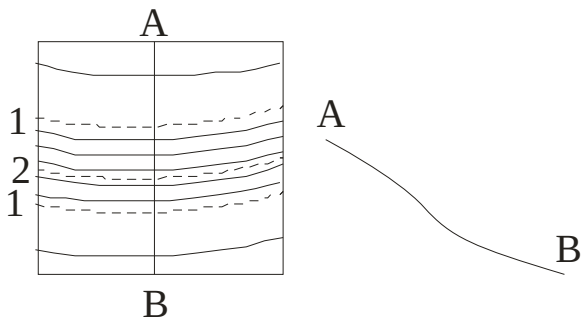


Рис. 8.6. Структурные линии на склоне

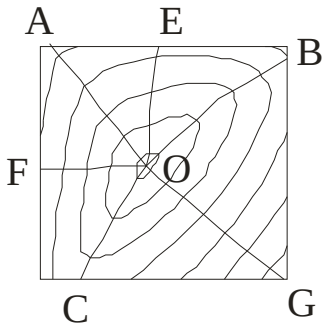


Рис. 8.7. Структурные линии на вершине

Поэтому в дальнейшем будем различать структурные линии первого и второго типов. Можно показать, что структурные линии разных типов должны чередоваться, как это имеет место на рис. 8.7. Структурные линии второго типа из-за особенностей человеческого восприятия менее заметны, чем линии первого типа. Предельным случаем структурных линий второго типа служат кривые, в каждой точке которых кривизна k_q сечения N_q равна нулю. Пример такой структурной линии приведен на рис. 8.8.

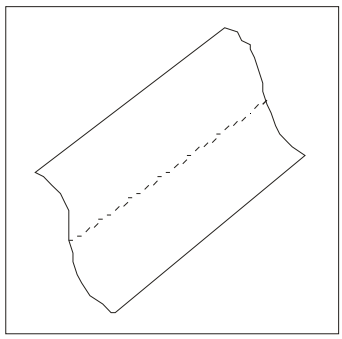


Рис. 8.8. Структурная линия второго типа

К структурным линиям следует относить также линии разрыва гладкости топографической поверхности вне зависимости от ее естественного или искусственного происхождения: верхние кромки обрывов, контуры оврагов, промоин, верхние и нижние края откосов и т. п. Линия разрыва гладкости поверхности – это предельный случай структурной линии первого типа, когда в

каждой ее точке касательная плоскость к поверхности изменяет свое положение скачком, а радиус кривизны $r_q = \frac{1}{k_q}$ главного нормального сечения N_q равен нулю.

В узлах число структурных линий четно и число структурных линий первого типа равно числу структурных линий второго типа. Вообще же, любую точку структурной линии можно рассматривать как узел, поскольку в каждой неомбилической точке пересекаются две линии кривизны. Структурные линии на гладкой поверхности не всегда ортогональны горизонталям. Но с течением времени под влиянием эндогенных сил топографическая поверхность, вероятно, стремится приобрести такую форму, что структурные линии первого порядка становятся ортогональными горизонталям на ней. Это свойство может быть использовано при построении информационных моделей топографических поверхностей.

Структурными точками на поверхности будем называть изолированные омбилические точки, точки пересечения или разветвления структурных линий (примеры даны на рис. 8.9 и 8.10), точки перегиба или излома структурной линии и точки локальных экстремумов.

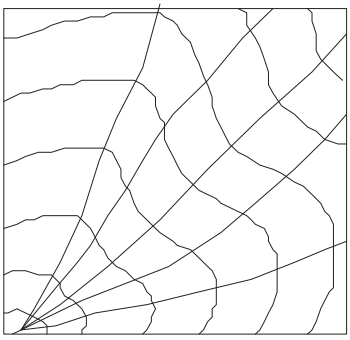


Рис. 8.9. Ветвление структурных линий

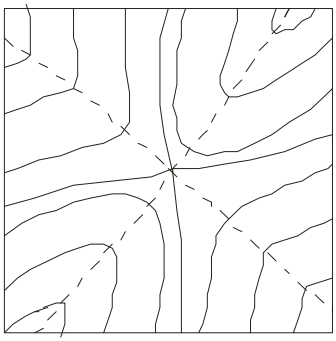


Рис. 8.10. Пересечение структурных линий

Из омбилических точек только изолированные мы считаем структурными в связи с тем, что существуют поверхности, состоящие только из омбилических точек (плоскость и сфера) и не имеющие структурных линий. В отличие от них, на поверхности эллипсоида вращения имеется две омбилические точки (полюсы) и одна структурная линия (экватор). Экватор может быть структурной линией первого или второго типа в зависимости от того, вокруг какой оси осуществляется вращение эллипса. Поверхность тора не имеет ни одной омбилической точки и ни одной структурной линии.

Точки локальных экстремумов отнесены к структурным не только в силу традиции. Они могут совпадать с изолированными омбилическими точками или с некоторыми точками пересечения структурных линий (см. рис. 8.7), но не обязательно. Их важное свойство, которое может быть использовано для повышения точности создаваемой модели топографической поверхности, – это равенство нулю производной по любому направлению, когда касательная

плоскость в точках локальных экстремумов занимает горизонтальное положение.

Совершенно справедливо утверждение, что структурные линии рельефа «облегчают распознавание отдельных форм его на местности и изображение их на карте и плане» [28, с. 775]. Более того, вся совокупность структурных линий и точек фактически формирует индивидуальный облик поверхности. Сходство двух участков поверхности определяется преимущественно сходством их структурных линий. Поэтому принято говорить, что структурные линии образуют скелет, или каркас поверхности. По этой причине структурные линии называют также скелетными и характерными. Точки локальных экстремумов в этом отношении имеют меньшее значение.

Чтобы подтвердить высказанное соображение, представим некоторый участок поверхности S, имеющий локальный минимум или максимум. Будем теперь изменять некоторым непрерывным образом ориентацию этого участка поверхности относительно системы пространственных координат XYZ. Очевидно, что точка локального экстремума при этом будет блуждать по поверхности. Точнее, экстремальными в разные моменты времени будут различные точки поверхности S, в совокупности образующие на ней некоторую траекторию. Положение же структурных линий и структурных точек (за исключением точки экстремума) по отношению к другим точкам поверхности S изменяться не будет. Объяснение этого состоит в том, что структурные линии и точки связаны с внутренней геометрией поверхности, тогда как экстремальность точек, как и положение изолиний, – следствие выбора системы координат, то есть акта, внешнего по отношению к поверхности S и, в известной мере, «случайного». Замечание о том, что точки локальных экстремумов не являются столь показательными для поверхности, как другие структурные линии и точки, следует понимать только таким образом.

Однако можно предположить, что с течением времени реальная топографическая поверхность в результате внешних воздействий приобретает такие формы, что точки локальных экстремумов либо совпадают с точками пересечения структурных линий, либо приближаются к ним на такое расстояние, что их можно считать совпадающими (см. рис. 8.7).

Описывая структурные линии, мы рассматриваем в их точках лишь одно главное нормальное сечение N_q, то есть «поперечное». Второе главное нормальное сечение N_p в каждой точке структурной линии является «продольным». Сечение N_q отличается тем, что его кривизна в точке структурной линии достигает экстремального значения. Кривизна же сечения N_p изменяется вдоль структурной линии некоторым произвольным и, возможно, не всегда непрерывным образом. Если в некоторой точке структурной линии кривизна k_p также экстремальна, то это служит признаком

того, что в данной точке структурная линия пересекается с другой структурной линией (или линиями).

На рис. 8.11, а представлен внешний вид гипотетической поверхности, 8.11, б – ее изображение горизонталями, 8.11, в – структурные линии на ней, а 8.11, д – профиль по замкнутой кривой, например, по окружности. Данный пример в гипертрофированном виде показывает, что существуют точки, в которых может быть более двух главных направлений. Подобного рода точки, из которых исходят, например, шесть структурных линий, встречаются иногда в местах ответвления отрогов горных хребтов.

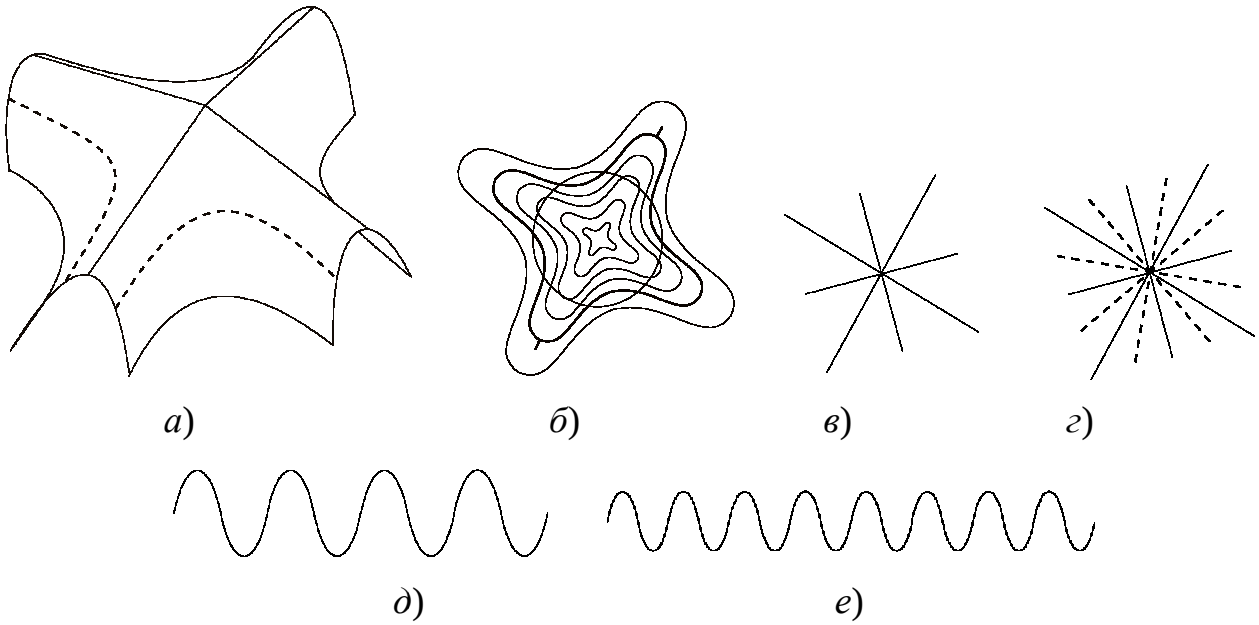
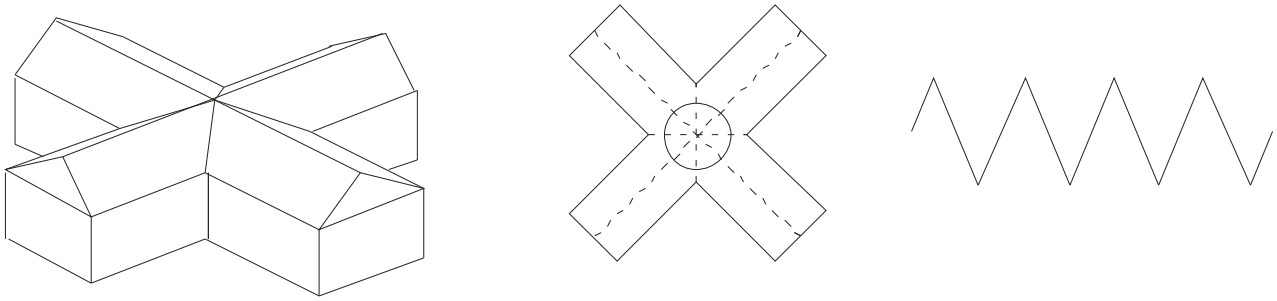


Рис. 8.11. Гипотетическая поверхность

Структурные линии гипотетической поверхности напоминают структурные линии на поверхности крыши здания, изображенные на рис. 8.12, б в виде штриховых линий. Скаты крыши являются плоскостями, следовательно, каждая их точка является омбилической. Поверхность же на рис. 8.11 не имеет омбилических точек. Профиль на рис. 8.11, д можно понимать как значения высот поверхности. Но если мы построим график значений кривизны нормальных сечений по той же окружности, то он будет иметь точно такой же вид.

Вообще-то, ответ на вопрос о том, сколько структурных линий исходит из вершины гипотетической поверхности на рис. 8.11, зависит от того, как мы понимаем кривизну нормального сечения. Если мы допускаем, что кривизна может иметь знак, то из вершины исходят восемь структурных линий (или в ней пересекаются четыре структурные линии). Но если мы под кривизной понимаем ее абсолютное значение, то количество структурных линий на рис. 8.11, в должно быть удвоено. Тогда к существующим структурным линиям следует добавить линии нулевой кривизны, исходящие из вершины, и мы получим рис. 8.11, г. В этом случае график значений кривизны также изменится и примет вид, показанный на рис. 8.11, е.



а) *б*) *в*)

Рис. 8.12. Поверхность крыши

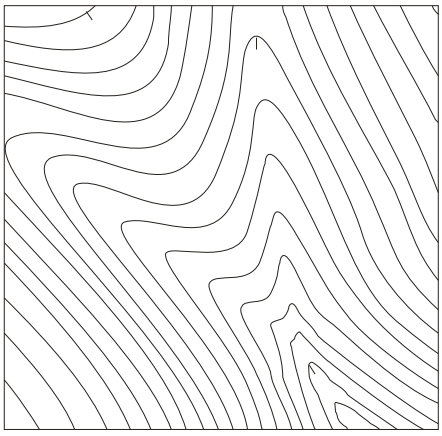


Рис. 8.13. Слияние тальвегов

Поверхности, изображенные на рис. 8.11, встречаются крайне редко, но участки, подобные показанному на рис. 8.13, на земной поверхности встречаются очень часто.

Понимание сущности структурных линий требуется для правильного выполнения съемки топографической поверхности и представления первичных (исходных) данных о топографических поверхностях, особенно при их информационном моделировании и автоматизированном создании карт и планов. Кроме того, оно необходимо при определении свойств топографической поверхности по ее информационной модели.

8.3. Исходные данные

Любая поверхность может быть задана либо своим уравнением, либо дискретным множеством точек. После получения исходных данных о топографической поверхности последняя для нас «перестает существовать», и топографическая поверхность такова, каковы исходные данные о ней. Таким образом, исходные данные о топографической поверхности уже являются ее некоторой моделью. Но эта модель, как правило, плохо структурирована, в связи с чем и возникает задача моделирования. Следовательно, моделирование может начинаться с процесса сбора данных о земной поверхности, то есть задолго до обработки на компьютере, как это имеет место при топометрическом методе сбора данных.

Получаемые при съемке исходные данные для создания информационных моделей топографических поверхностей представляют собой совокупность (образов) конечного числа точек, отобранных в соответствии с некоторым критерием или условием выборки из бесконечного множества точек моделируемой поверхности. В качестве критерия выборки может использоваться принадлежность точки:

– к структурным точкам и линиям поверхности; данный критерий используется в подавляющем большинстве случаев при наземных и стереотопографических съемках;

– к профилям (или галсам); такой подход используется при крупномасштабной наземной съемке залесенной местности и при съемке дна водоемов и акваторий;

– к горизонталям; данный способ применяется при создании информационных моделей топографической поверхности картометрическим методом и при стереотопографической съемке;

– к узлам регулярной сетки, данный критерий применим при любом методе съемки, но крайне неэффективен.

На практике часто используется некоторая комбинация этих критериев. Выбор критерия выборки диктуется такими факторами, как технология сбора исходной информации, используемое оборудование, сложность поверхности, квалификация исполнителей, имеющиеся методы моделирования, требования к точности, стоимости, времени создания моделей и рядом других, вплоть до предпочтений исполнителей.

Если различные критерии рассматривать с точки зрения эффективности методов сбора, то критерий принадлежности к характерным точкам и линиям топографической поверхности находится вне конкуренции. Структурные линии и точки являются наиболее презентабельными, наиболее информативными. Использование данного критерия позволяет представить топографическую поверхность минимумом исходных точек при соблюдении заданной точности, что особенно важно при топометрическом методе сбора данных, то есть при измерениях непосредственно на земной поверхности.

С позиций метода моделирования важно то, что конкретному условию выборки соответствует определенная схема выборки – взаимное положение точек в плане. Схемы выборки можно подразделять на регулярные, полурегулярные и нерегулярные (рис. 8.14).

В регулярных схемах выборки в результате соединения соседних точек прямыми на плоскости образуется периодически повторяющийся рисунок (орнамент), покрывающий всю область моделирования. Известно, что на плоскости может быть 17 типов орнамента, но, по понятным причинам, на практике используются лишь простейшие случаи: сетка равносторонних или прямоугольных треугольников, квадратов или шестиугольников (рис. 8.14, *а–д*).

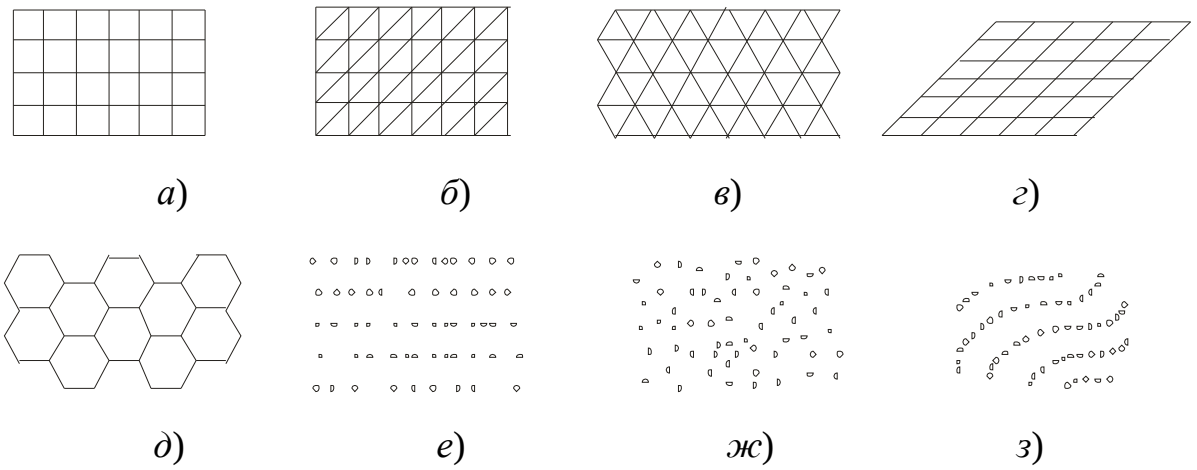


Рис. 8.14. Схемы выборки

Полурегулярные схемы выборки характеризуются тем свойством, что периодичность рисунка сохраняется лишь в одном направлении (рис. 8.14, е). Сюда можно отнести прямоугольную сетку с переменным шагом по каждой из ее сторон, сетку равнобедренных треугольников, профили. Некоторые полурегулярные схемы выборки при помощи топологического преобразования могут быть трансформированы в регулярные схемы.

В нерегулярных схемах выборки периодичность рисунка не сохраняется ни по одному направлению (рис. 8.14, ж, з). В общем случае нерегулярные схемы выборки не могут быть преобразованы в регулярные и образуются при съемке характерных (структурных) точек топографической поверхности или точек на горизонталях.

Исходные данные о топографической поверхности представляют собой перечень точек, заданных значениями своих координат и высот или глубин, и перечень структурных линий с их характеристиками. Структурные линии на однозначной поверхности представляют собой плоский граф. Если поверхность не является однозначной функцией двух переменных, то граф структурных линий уже не является плоским, но его можно сделать плоским, если деформировать поверхность.

8.4. Оценка сложности кривых и поверхностей

В процессе моделирования кривых и поверхностей естественным образом возникает вопрос об их сложности. Термины «сложная поверхность» или «сложная кривая» употребляются часто, но что за ними стоит, не очень ясно. Проблема оценки сложности поверхности имеет и чисто практическое значение, поскольку с ней сталкиваются, например, при расчете трудоемкости топографической съемки или стоимости создания информационных моделей геопространства.

В топографии и картографии оценку сложности поверхности принято осуществлять с помощью эталонов сложности. Эти эталоны представляют собой изображения рельефа разного типа (пойменного, равнинного, холмистого, горного, высокогорного и т. п.) с указанием для каждого типа соответствующей категории сложности (от 1 до 10). Трудоемкости двух соседних категорий отличаются на 15 % и более, а трудоемкость 10-й категории превышает трудоемкость 1-й категории более чем в 10 раз. Определение сложности конкретной поверхности производится ее визуальным сравнением с эталонами, и расхождения в оценке иногда составляют даже не одну, а две категории. По этой причине сложность выполненных работ нередко является предметом дискуссий между исполнителями и их руководителями, а иногда – причиной производственных конфликтов.

Поскольку совершенно неясно, что может служить мерой сложности кривых и поверхностей, постольку определение такой меры целесообразно начать с определения ее желательных, или «разумных», свойств. Очевидно, что мера сложности кривой должна обладать следующими свойствами, которые могут рассматриваться как аксиомы сложности кривых.

1. Мера сложности кривой должна быть независимой от размера (масштаба) кривой, то есть должно выполняться равенство

$$C(x, y) = C(kx, ky).$$

2. Мера сложности кривой должна быть независимой от ее ориентации, или иначе, должно выполняться равенство

$$C(x, y) = C(X, Y),$$

где

$$\left.\begin{aligned} X &= x\cos\alpha - y\sin\alpha \\ Y &= x\sin\alpha + y\cos\alpha \end{aligned}\right\}.$$

Первые два свойства означают, что сложность кривой не должна быть зависимой от выбора системы координат. Следовательно, сложность кривой относится к внутренней геометрии.

3. Сложность кривой или ее участка не может быть отрицательной величиной:

$$C \geq 0;$$

$$C_j \geq 0 \quad (j = 1, \ldots, n),$$

где C_j – сложность j-го участка кривой.

4. Сложность j-го участка кривой не может быть больше сложности C кривой в целом:

$$C_j \leq C.$$

5. Мера сложности должна обладать аддитивностью. Это означает, что при разбиении кривой на n участков сложность кривой в целом должна равняться сумме значений сложности ее участков, или иначе, должно выполняться равенство

$$C = \sum_{j=1}^{n} C_j,$$

где C_j – значение сложности j-го участка.

6. Мера сложности должна быть в равной степени применимой для оценки гладких и негладких кривых.

7. За нулевое значение сложности может быть принята сложность вырожденной кривой – точки, то есть

$$C(P) = 0.$$

Такое решение обусловлено тем, что кривая в таком случае отсутствует. Поэтому естественно считать, что ее сложность равна 0.

8. За единицу принимается значение сложности кривой L, представляющей собой отрезок прямой:

$$C(L) = 1.$$

Очевидно, что с математической точки зрения сложность отрезка прямой не зависит от его длины. Два любых прямолинейных отрезка являются «подобными», поэтому можно считать, что сложность одного из них равна сложности другого. Равенство значений их сложности следует из свойств 1 и 2, но вопрос о конкретном значении сложности прямолинейных отрезков при этом остается открытым. Поэтому данная аксиома является дополнением к аксиомам 1 и 2.

9. Сложность n кривых является суммой

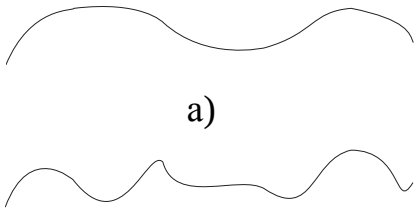
$$C = \sum_{k=1}^{n} C_k ,$$

где C_k есть сложность отдельной кривой.

После формулировки свойств меры сложности можно обратиться к выбору самой меры. На рис. 8.15 изображены две кривые. Интуитивно кривая b представляется более сложной, чем кривая a. Эта интуиция основывается на смутном ощущении того, что сложность кривой или поверхности каким-то образом связана с их кривизной. Следовательно, в качестве меры сложности кривой можно было бы избрать интеграл ее кривизны

$$C = \int_0^t \frac{|f''(t)|}{\left(1 + (f'(t))^2\right)^{3/2}} dt ,$$

где t – некоторый параметр, например, длина кривой.



а)

б)

Рис. 8.15. Кривые разной сложности

Но при таком решении возникает проблема оценки сложности негладких кривых, так как кривизна кривой в точке излома равна бесконечности. Чтобы избавиться от этого недостатка, можно считать отдельной кривой каждый гладкий кусок кривой. Но тогда возникает следующая проблема: если изображение представляет собой сколь угодно большое множество кривых, каждая из которых является прямолинейным отрезком, и интеграл кривизны каждой из которых равен нулю, то их суммарная сложность в таком случае также будет равна нулю, с чем нельзя согласиться. По этой причине, а также в соответствии с аксиомой 8 можно предположить, что сложность кривой может быть выражена формулой

$$C = 1 + \int_0^t \frac{|f''(t)|}{\left(1 + (f'(t))^2\right)^{3/2}} dt .$$

Но при выполнении топографических или картографических работ аналитическое представление кривых неизвестно. Поэтому необходимо найти некоторое приближение к интегралу кривизны либо замену ему. Кроме того, точнее говорить не о сложности некоторой кривой, а о сложности ее представления.

В качестве эталона для оценки сложности кривой подсознательно используется образ прямой. Мы готовы признать, что кривая тем сложнее, чем больше она отличается от прямой. Поэтому первая мысль, которая возникает при попытке выразить сложность кривой с помощью числа, связана с отклонением от прямой (рис. 8.16). Первоначально при моделировании мы можем заменить кривую прямой AB или ломаной линией. Максимальное отклонение Δ кривой от ломаной можно расценивать как приращение сложности кривой (или ее представления). Мы можем последовательно заменять каждый отрезок ломаной двумя отрезками (рис. 8.16, б), считая, что сложность представления кривой на каждом шаге увеличивается на соответствующее отклонение Δ_i.

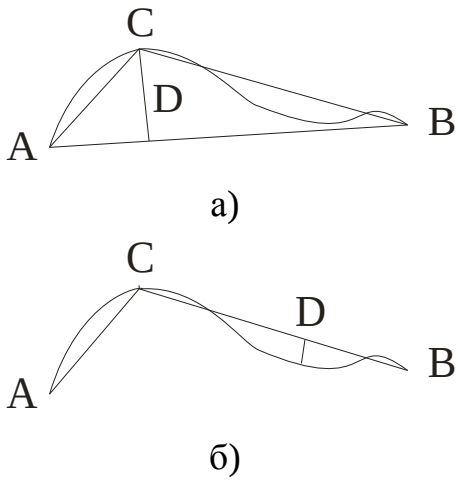


Рис. 8.16. Отклонение от прямой

Но такое решение представляется сомнительным, поскольку приходится признать, что отклонение Δ характеризует в большей степени точность представления кривой, нежели ее сложность.

Пусть имеется некоторая разомкнутая плоская кривая AF (рис. 8.17), которую нужно представить в виде ломаной с точностью не хуже Δ, под которой будем понимать максимальное отклонение разомкнутой кривой от аппроксимирующей ломаной на всем ее протяжении. В первом приближении мы можем кривую заменить замыкающей AF. Изменим масштаб (систему координат) таким образом, чтобы длина замыкающей AF была равной 1. Для этого достаточно изменить прямоугольные координаты X, Y по формулам

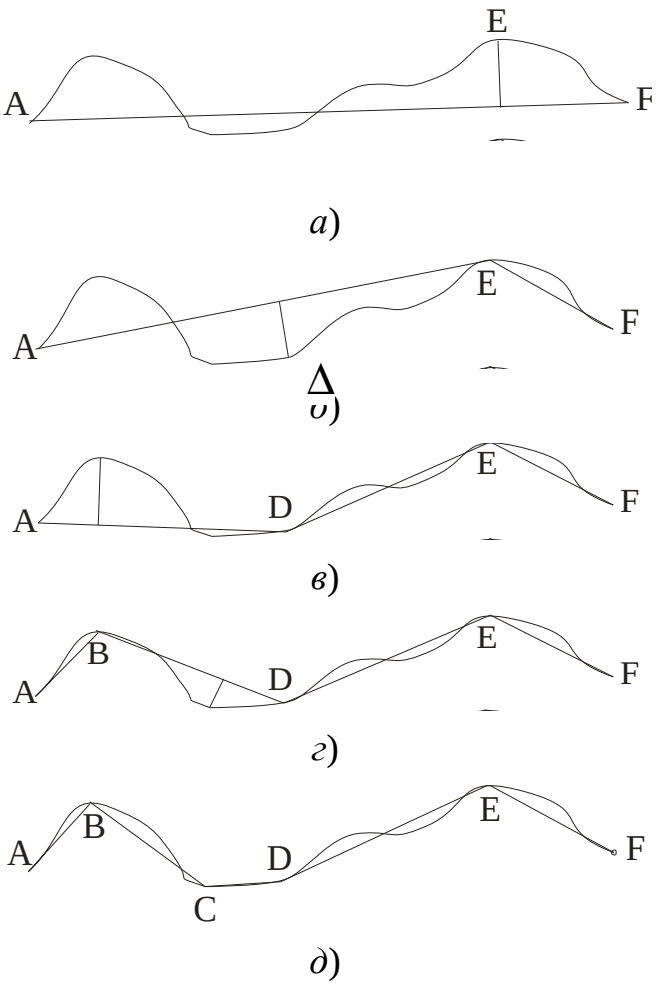


Рис. 8.17. К определению сложности прямой

$$\left.\begin{aligned} x &= \frac{X}{L} \\ y &= \frac{Y}{L} \end{aligned}\right\},$$

где $L = \sqrt{(X_n - X_1)^2 + (Y_n - Y_1)^2}$ – длина замыкающей AF.

Нормируя замыкающую AF, мы получаем возможность сравнивать кривые разной длины. Но это будет справедливо только тогда, когда отклонение Δ мы также нормируем, то есть выразим в единицах расстояния между первой и последней точками кривой:

$$\delta = \frac{\Delta}{L},$$

где δ – относительная ошибка представления кривой. С практической точки зрения нас интересует не столько сложность кривой сама по себе, сколько сложность кривой при необходимой точности ее представления. Понятие относительной ошибки δ дает возможность сравнивать сложность кривых различной длины при заданной абсолютной или относительной точности их представления.

Предположим, что точность представления на рис. 8.17, а неудовлетворительна, и наибольшее отклонение от замыкающей AF кривая имеет
в точке E. Поэтому заменим отрезок прямой AF ломаной AEF (рис. 8.17, б). Сложность C_{AEF} ломаной AEF выше сложности C_{AF} отрезка AF и может быть представлена как сумма C_{AF} и некоторого приращения

$$C_{AEF} = C_{AF} + c_{AEF}. \qquad (8.6)$$

Приращение сложности c_{ijk} (по отношению к сложности C_{ik}) при замене произвольного отрезка IK на ломаную IJK представим в виде выражения

$$c_{ijk} = C_{ij} + C_{jk} - C_{ik}. \qquad (8.7)$$

Тогда сложность двух отрезков IJ и JK будет равна

$$C_{ijk} = C_{ik} + (C_{ij} + C_{jk} - C_{ik}) = C_{ij} + C_{jk}.$$

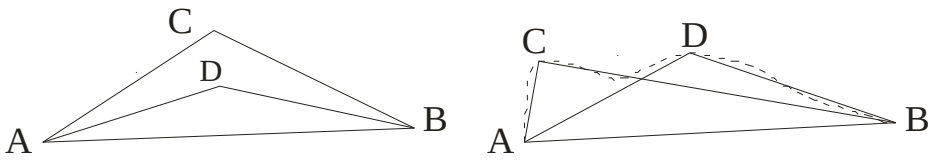
В качестве значения сложности произвольного отрезка IJ ломаной, представляющей кривую, примем нормированную длину этого отрезка

$$C_{ij} = l_{ij} = \frac{L_{ij}}{L}.$$

Следовательно, увеличение сложности будет равно приращению относительной длины ломаной при замене одного отрезка двумя другими

$$c_{ijk} = l_{ij} + l_{jk} - l_{ik}.$$

Чтобы объяснить такое решение, рассмотрим два примера. На рис. 8.18, а сложность ломаной ACB интуитивно представляется выше, чем ломаной ADB, поскольку точка C отстоит от отрезка AB на большее расстояние, чем точка D. Следовательно, точка C более «информативна» или более оригинальна.



а) б)

Рис. 8.18. Ломаные разной сложности

На рис. 8.18, б точки C и D удалены от AB на одинаковое расстояние, но ломаная ACB более изломана, и поэтому она «сложнее». Слева и справа на рис. 8.18 ломаные ACD более искривлены и имеют большую длину, чем ADB. Естественно принять, что чем больше искривлена кривая, тем выше ее сложность. Но чем выше кривизна кривой, тем больше ее длина. Нормированная длина ломаной позволяет косвенным образом судить о сложности аппроксимируемой кривой. Поэтому есть основания считать, что увеличение сложности представления кривой может быть выражено относительным приращением длины при замене отрезка двумя отрезками (см. рис. 8.18).

Для проверки нашего решения рассмотрим случай, когда на прямолинейном отрезке IJ мы ставим дополнительную точку K (рис. 8.19). Получаемое при этом приращение сложности будет равно

$$c_{ijk} = l_{ik} + l_{kj} - l_{ij}. \qquad (8.8)$$

Но $l_{ik} + l_{kj} = l_{ij}$, поэтому $c_{ijk} = 0$. Таким образом, наше решение не противоречит здравому смыслу.

Вернемся к представлению разомкнутой кривой на рис. 8.17. Если точность ее аппроксимации по-прежнему не достаточна, то мы последовательно заменяем каждый отрезок ломаной двумя, пока не достигнем заданной абсолютной или относительной точности представления кривой (рис. 8.17, б–д). Оценка сложности кривой на каждом таком шаге производится по формуле

Рис. 8.19. Вставка в отрезок

$$C_i = C_{i-1} + c_i, \qquad (8.9)$$

где C_{i-1} и C_i обозначают соответственно предыдущее и текущее значение сложности представления кривой, а c_i – приращение сложности, вычисляемой по формуле (8.7).

С добавлением точек на прямолинейном отрезке в представлении кривой ничего не изменяется, поэтому можно считать, что количество информации о кривой при этом также сохраняется, а увеличивается только при добавлении

точек, не принадлежащих аппроксимирующей ломаной. При этом также можно считать, что количество информации тем больше, чем больше это отклонение.

Пусть дана замкнутая кривая и две точки на ней A и B такие, что AB – наибольший диаметр, то есть максимальное расстояние между любыми двумя точками кривой (рис. 8.20). Тогда задача оценки сложности замкнутой кривой сводится к задаче оценки сложности двух разомкнутых кривых L_1 и L_2. Значение сложности замкнутой кривой может вычисляться по формуле

$$C_{12} = C_1 + c_{12}, \qquad (8.10)$$

где C_{12} – сложность замкнутой кривой; C_1 – сложность кривой L_1; а c_{12} – приращение сложности при добавлении кривой L_2 к L_1. Но значение приращения сложности равно

$$c_{12} = C_2,$$

поэтому сложность замкнутой кривой может быть представлена выражением

$$C_{12} = C_1 + C_2. \qquad (8.11)$$

Эта же формула может быть использована для вычисления сложности суммы (объединения) двух кривых.

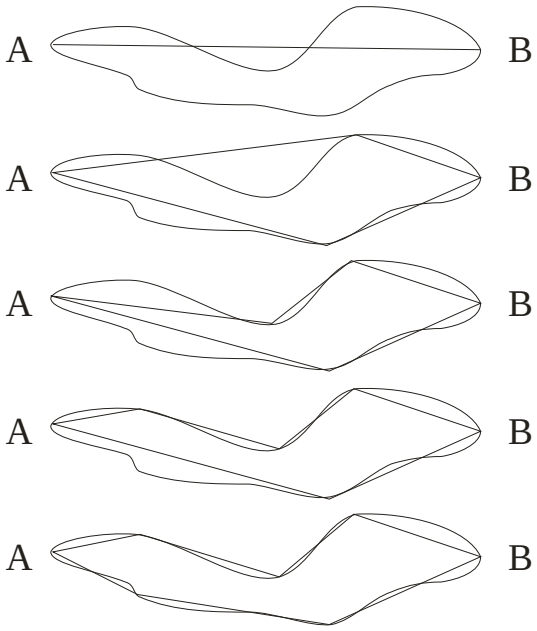


Рис. 8.20. К оценке сложности замкнутой кривой

Чтобы уменьшить ошибку Δ представления кривой в виде ломаной, мы можем увеличивать число n звеньев ломаной. Тогда при $n \to \infty$ будет $\Delta \to 0$, и сложность кривой может быть представлена выражением

$$C = \lim_{n\to\infty} \sum_{i=1}^{n} c_i = l, \qquad (8.12)$$

где l – нормированная длина кривой.

Сложность кривой может быть сопоставлена с информацией о кривой, если придерживаться разнообразностной трактовки информации, когда количество информации служит мерой многообразия, сложности. Очевидно, что количество информации, содержащееся в ломаной, аппроксимирующей кривую, не может быть больше, чем количество информации в самой кривой. Если с этой точки зрения посмотреть на оценку сложности ломаных, то можно согласиться с тем, что возможна трактовка сложности ломаной как количества информации о представлении аппроксимируемой кривой.

В частности, если мы имеем одну точку, и больше о кривой ничего не известно, то мы имеем полную неопределенность, и количество информации можно считать равным нулю. В случае ограниченного участка (отрезка) прямой

мы имеем полную определенность, и количество информации можно принять равным 1.

Чтобы облегчить решение задачи оценки сложности поверхностей, будем рассматривать только поверхности, не обязательно выпуклые, гомеоморфные сфере. Каждая замкнутая выпуклая поверхность является гомеоморфной сфере. Если выпуклая поверхность не замкнута, то она гомеоморфна поверхности с краем (ограниченному участку сферы или плоскости). Но не каждая невыпуклая поверхность может быть гомеоморфной сфере или ее участку (например, тор). Далее будем предполагать, что поверхность хотя и невыпуклая, но гомеоморфна сфере или ее участку. Такое решение дает возможность вписать тем или иным образом сферу в поверхность, ввести на сфере систему криволинейных координат (u, v) и выражать пространственное положение произвольной точки поверхности через криволинейные координаты и высоту точки – расстояние от данной точки до сферы.

Подход к оценке сложности представления кривых может быть распространен на оценку сложности поверхностей. Произвольный участок поверхности в первом приближении может быть представлен плоскостью, проходящей через три точки, принадлежащие поверхности и не лежащие на одной прямой. Поэтому в качестве простейшей геометрической фигуры, представляющей поверхность, выберем треугольник. При добавлении новой точки к трем имеющимся (к вершинам треугольника) поверхность и ее сложность могут изменяться. При этом возможны следующие случаи (рис. 8.21).

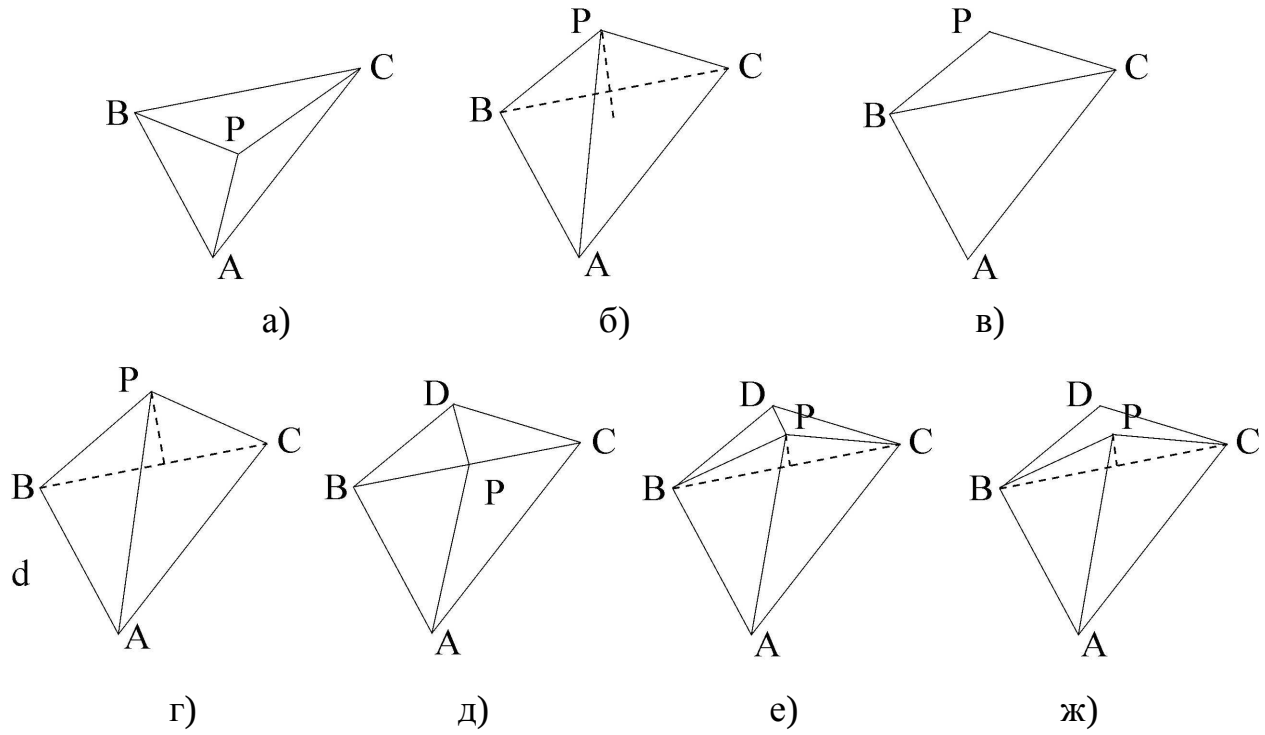


Рис. 8.21. К приращению сложности поверхности

Если к трем вершинам треугольника ABC добавляется точка P, лежащая в плоскости треугольника (рис. 8.21, а), то форма поверхности, а, следовательно,

и ее сложность не изменяются. Площадь поверхности при этом также не изменяется.

Если добавляемая точка не принадлежит плоскости треугольника, но находится внутри его области определения (рис. 8.21, б), то треугольник ABC заменяется тремя новыми. Форма поверхности изменяется, следовательно, изменяется и сложность поверхности.

Если точка P лежит вне области треугольника ABC (рис. 8.21, в), то поверхность изменяется независимо от того, лежит точка P в плоскости, проходящей через вершины A, B, C, или нет. При этом к уже существующему треугольнику может добавляться новый, как это показано на рис. 8.21, в, либо треугольник ABC уничтожается и создаются два новых: ABP и ACP, что на рис. 8.21 не показано. Форма участка поверхности и его площадь в обоих случаях изменяются, поэтому можно считать, что при этом изменяется и сложность поверхности.

Добавляемая точка P может попасть на ребро, и это ребро является граничным, то есть принадлежит только одному треугольнику (рис. 8.21, г). Если точка P не принадлежит отрезку прямой BC, то треугольник ABC заменяется двумя треугольниками. Форма поверхности и ее площадь при этом изменяются. Следовательно, сложность представления поверхности также изменяется. Если же точка P принадлежит отрезку прямой BC, то поверхность и ее площадь при таком добавлении сохраняются, и можно считать, что сохраняется и значение сложности поверхности.

Пусть добавляемая точка P принадлежит стороне, являющейся общей для двух треугольников (рис. 8.21, д). Форма поверхности не изменяется, и у нас нет оснований считать, что сложность поверхности при этом изменилась.

Если проекция точки P принадлежит проекции ребра, но сама точка ребру не принадлежит (рис. 8.21, е, ж), то форма поверхности изменяется. При этом возможны два варианта образования новых треугольников: либо два существовавших треугольника удаляются и создаются четыре новых, как показано на рис. 8.21, е, либо треугольник BCD сохраняется, а вместо треугольника ABC создаются три новых, один из которых лежит в вертикальной плоскости (рис. 8.21, ж). В последнем случае поверхность уже не будет выпуклой и однозначной.

Во всех перечисленных случаях при более точном представлении поверхности путем добавления точек одновременно с усложнением поверхности происходит увеличение ее площади. Поэтому можно считать, что приращение сложности поверхности связано с увеличением ее площади

$$C_i = C_{i-1} + c_i, \qquad (8.13)$$

где C_{i-1} – сложность поверхности до добавления i-й точки; C_i – сложность поверхности, представленной i точками; c_i – приращение сложности при добавлении i-й точки.

Представлением поверхности при этом является многогранник, не обязательно выпуклый, каждая грань которого – треугольник. Давая оценку

сложности кривых – одномерных геометрических объектов – мы выбрали в качестве меры их сложности нормированную длину. Если придерживаться последовательности, то при оценке сложности поверхностей – двумерных объектов – в качестве меры сложности следует использовать нормированную площадь поверхности.

За единицу сложности примем сложность проекции моделируемой поверхности на «поверхность относимости», то есть площадь области определения функции, описывающей представляемую (оцениваемую) поверхность. Если поверхность относимости является плоскостью, то представляется разумным сравнивать сложность любой поверхности со сложностью ограниченного куска плоскости. Сложность последнего может быть принята за единицу.

Если поверхность относимости является сферой или ее участком, то за единицу сложности следует принять сложность сферы (или ее участка). Плоскость и сфера являются простейшими поверхностями и обладают тем общим свойством, что их кривизна постоянна.

Вне зависимости от того, что собой представляет «поверхность относимости», плоскость или сферу, можно согласиться считать единицей сложности площадь области определения функции, описывающей поверхность. Тогда сложность поверхности будем оценивать следующим образом.

Для простоты будем полагать, что исследуемая ограниченная (имеющая конечную площадь) поверхность S задана в трехмерном евклидовом пространстве и может быть представлена функцией $z = f(x, y)$, имеющей область определения D. Пусть на поверхности S задана некоторая триангуляция, проекция которой на координатную плоскость xy с достаточной точностью представляет область определения D.

Обозначим площадь проекции триангуляции S_0. Сложность C_i каждого пространственного треугольника на поверхности представим как отношение его площади S_i к площади S_0

$$C_i = \frac{S_i}{S_0}. \qquad (8.14)$$

Сложность C всей поверхности в целом будет равна

$$C = \frac{1}{S_0} \sum_{i=1}^{n} S_i. \qquad (8.15)$$

Сложность нескольких поверхностей может быть получена суммированием значений сложности для каждой отдельной поверхности. Поскольку сложность и ее приращение величины относительные, постольку не столь принципиальным является вопрос о выборе поверхности, значение сложности которой принимается равным 1.

Возможно, что предложенная оценка сложности кривых и поверхностей не отличается глубиной и может показаться слишком простой для того, чтобы быть

правильной. Слабость предложенной трактовки в том, что сложность поверхности зависит от выбора поверхности относимости.

Сложность замкнутой поверхности может приниматься равной сумме двух незамкнутых поверхностей, то есть поверхностей с краем. Так, сложность всей земной поверхности равна сумме сложности поверхности ее северного и южного полушарий.

Альтернативными вариантами меры сложности поверхностей, как и кривых, могло бы служить использование интеграла кривизны, что сложнее, или числа точек, представляющих кривую или поверхность, что еще проще.

Число точек, представляющих кривую или поверхность, в качестве меры их сложности также не способно служить критерием сложности их формы. Но возможно, что число точек может использоваться для оценки структурной сложности представления кривых и поверхностей, если такое представление оценивать как сложность графа.

8.5. Математические модели топографической поверхности

Давая определение математической модели топографической поверхности, будем исходить из общего определения математических моделей.

С математической точки зрения, физическая поверхность Земли представляет собой бесконечное точечное множество. Более того, окружающее нас реальное физическое пространство с достаточной степенью точности традиционно описывается такой математической структурой, как трехмерное евклидово (точечно-векторное) пространство. Это соглашение хорошо соответствует нашим интуитивным представлениям о свойствах физического пространства и столь часто используется, что о существовании самого соглашения обычно забывают.

Признав адекватность математического евклидового пространства пространству физическому, мы тем самым наделяем второе свойствами первого. Тогда изучение свойств реального пространства может сводиться к изучению свойств математического пространства. Свойства и отношения между элементами евклидового пространства рассматриваются в курсах линейной алгебры, и останавливаться на них здесь нет необходимости. Но далее следует помнить, что в определении математической модели топографической поверхности элементы базового множества являются точками евклидового пространства и обладают всеми соответствующими свойствами и отношениями.

Задача математического моделирования (физической) топографической поверхности F заключается в получении некоторой мыслимой поверхности H, в том или ином смысле близкой к F. Эту задачу можно также определить как замену точечного множества F точечным множеством H. В отличие от F, представляющего собой поверхность, точечное множество H, вообще говоря, может не быть поверхностью, а представлять собой дискретное множество точек. В любом случае множество H должно быть близким в некотором смысле к поверхности F и может быть использовано для конструирования по дискретному точечному множеству тем или иным способом поверхности, опять же близкой к F.

Абстрактным образом точки топографической поверхности может служить упорядоченная тройка чисел (X, Y, Z) – прямоугольных координат в трехмерном евклидовом пространстве. Однако, как уже говорилось выше, для удобства вводится поверхность относимости с заданной на ней системой координат Oxy, и топографическая поверхность трактуется как некоторая гипотетическая функция двух переменных $f(x, y)$, где x, y могут быть прямоугольными координатами на плоскости, геодезическими или геоцентрическими координатами на эллипсоиде и т. п. Поверхность при этом может описываться выражениями вида (8.3)–(8.5).

Представления о топографической поверхности как о непрерывной в физическом смысле дают основания считать ее непрерывной и с математической точки зрения: достаточно близкие точки (x_1, y_1) и (x_2, y_2) должны иметь близкие значения высот $f(x_1, y_1)$ и $f(x_2, y_2)$. Следовательно, функцию $f(x, y)$ можно считать непрерывной. К сожалению, функция $f(x, y)$ не является аналитической (или регулярной, голоморфной, дифференцируемой), поскольку участки реальной топографической поверхности часто не являются гладкими (обрывы, овраги,...). Кроме того, функция $f(x, y)$ может быть неоднозначной (например, нависающие скалы), но поскольку трудности, возникающие при моделировании неоднозначных топографических поверхностей на ЭВМ значительны, а необходимость в их представлении возникает редко, то, видимо, такие попытки вообще не предпринимались. Поэтому будем считать функции, описывающие топографические поверхности, однозначными, если не указано обратное.

Область определения D функции $f = (x, y)$ связана с поверхностью относимости E соотношением $D \subseteq E$. Множество точек E, образующих поверхность относимости, является бесконечным, ограниченным, связным и замкнутым. Другими словами, множество E представляет собой замкнутую область в евклидовом пространстве, например, сферу или эллипсоид. Такими же свойствами обладает и точечное множество F – множество всех точек физической поверхности Земли. Если объектом моделирования является вся земная поверхность, понимаемая как поверхность суши и дна океанов, то $D = E$.

Если земная поверхность трактуется только как поверхность суши или ее некоторый ограниченный участок $(D \subset E)$, то она обладает уже несколько отличными свойствами. Множество точек F топографической поверхности при этом также бесконечно и ограничено, но не обязательно связно. Тогда F может быть представлено как объединение конечного числа связных подобластей

$$F = F_1 \cup F_2 \cup ... \cup F_n,$$

не имеющих общих границ. Произвольная подобласть F_i (i = 1, 2, …, n) является ограниченной, замкнутой и может быть либо односвязной, либо

многосвязной. Граница любой области – некоторая, не обязательно гладкая, замкнутая пространственная кривая.

Вне зависимости от области определения функция может быть представлена конечным или бесконечным множеством точек. Бесконечные множества задаются описанием. Говорят, что множество H является образом множества D или что H получено из D:

$$H = \{h(x, y) \mid (x, y) \in D\},$$

где D – область определения отображения.

Конечные множества указываются перечислением, что равносильно выделению подмножества с помощью предиката:

$$H = \{h \mid P(h)\}.$$

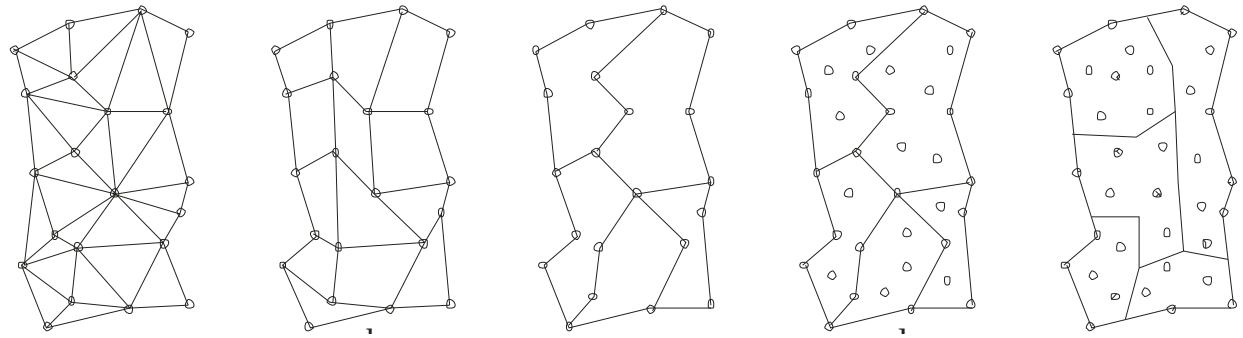
Выделение H с помощью предиката может осуществляться двумя способами. При использовании первого способа перечисляются все элементы декартова произведения $D \times Z$ (где D – область определения, а Z – область значений), и те из них, которые принадлежат $H \subset D \times Z$, отмечаются каким-либо образом. При использовании второго способа перечисляются только те элементы декартова произведения $D \times Z$, для которых значение предиката истинно. Хотя оба способа обладают как преимуществами, так и недостатками, на практике чаще используется второй, как более компактный.

Таким образом, математические модели топографических поверхностей прежде всего различаются по типу базового множества H. Если точечное множество H является бесконечным, топографическая поверхность представляется как совокупность всех точек, удовлетворяющих некоторому уравнению $h = h(x, y)$, и математическая модель называется аналитической, или непрерывной. При этом требуется, чтобы функция $h = h(x, y)$ была близка к физической поверхности $f(x, y)$. Возможность воспроизведения гладких функций с любой наперед заданной точностью при помощи рядов известна из математического анализа.

Первой альтернативой непрерывным моделям служат кусочно-непрерывные модели. Для построения кусочно-непрерывной модели поверхности область определения функции (область моделирования) разделяется на подобласти (или элементы) так, что подобласти образуют либо разбиение, либо покрытие всей области. Разбиение области D на m подобластей означает, что $D = D_1 \cup \ldots \cup D_m$ и $D_i \cap D_j = \varnothing$, если $i \neq j$. Иными словами, при разбиении подобласти покрывают всю область моделирования и являются непересекающимися. Покрытие области моделирования понимается как такое ее деление на подобласти, когда подобласти покрывают всю область моделирования, но пересечение двух соседних подобластей не пусто, то есть имеет место $D_i \cap D_j \neq \varnothing \quad (i \neq j)$. Но иногда такое строгое разделение не проводится, и разбиение называют также покрытием.

Разбиение обычно строится таким образом, что границы элементов представляют собой непересекающиеся многоугольники, вершины которых совпадают с исходными точками (рис. 8.22, а–г). Иногда разбиение области моделирования осуществлялось так, чтобы граница между различными участками не проходила через исходные точки (рис. 8.22, д). Но подобные методы могут применяться разве только от безысходности и отчаяния, поскольку возникает проблема стыковки соседних участков по высоте. Далее будем различать разбиение области моделирования на произвольные многоугольники, выпуклые четырехугольники и треугольники.

Если блоки достаточно крупные (по числу исходных точек), то их границы могут не связываться с положением опорных точек, хотя и эта мера не устраняет проблемы стыковки. Но идея разбиения поверхности на простые элементы приводит к необходимости их «закрепления» в исходных точках. Если исходные точки являются узлами регулярной сетки, то хранение границ элементов не требуется. При нерегулярной схеме выборки нет иного выхода, как описывать границы элементов перечислением.



а) *б*) *в*) *г*) *д*)

Рис. 8.22. Разбиение области моделирования

Четырехугольные элементы, опирающиеся на произвольно расположенные исходные точки, здесь рассматриваться не будут, так как они часто не дают удовлетворительного разбиения поверхности.

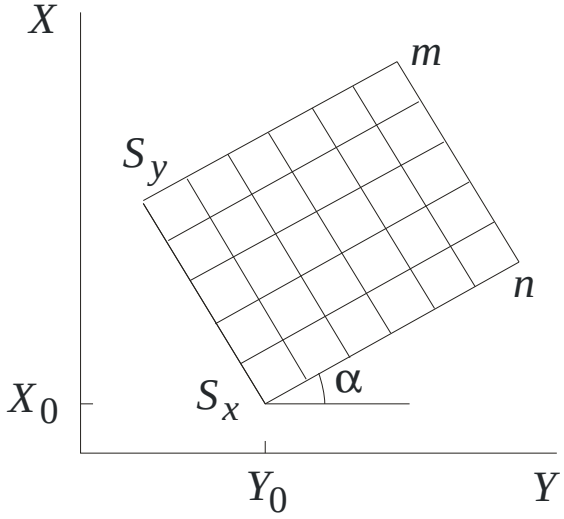


Рис. 8.23. Регулярная модель

Разбиение участка на блоки или фрагменты и кусочное представление поверхности позволяют повысить точность моделирования внутри блоков, но требуют стыковки блоков и запоминания их границ. Чем сложнее границы блоков, тем сложнее обеспечить непрерывность функции и ее производных, и тем больше требуется дополнительной памяти при описании границ. С уменьшением размеров блоков их конфигурация упрощается, уменьшается число точек в блоке, и, как следствие, понижается

порядок разложения функции в ряд, возникают предпосылки для повышения точности и стыковки блоков. Перечисленные обстоятельства имеют столь важное значение, что основное развитие методов моделирования осуществляется в этом направлении, то есть в применении кусочно-непрерывных моделей. В пределе моделируемая поверхность считается составленной из простейших элементов, каждый из которых описывается своим уравнением.

Если базовое множество конечно, то есть поверхность представляется набором дискретных точек, то модель называется дискретной. Аналитические модели представляют собой непрерывные (всюду плотные) и, следовательно, бесконечные множества; дискретные модели являются конечными множествами.

Разбиение области моделирования на более простые элементы позволяет избавиться от необходимости решать большие системы уравнений, но порождает другие проблемы. Основная из них заключается в том, как обеспечить непрерывность функции и ее производных на границе двух смежных элементов (если их граница является линией разрыва гладкости, то требуется только непрерывность значений самой функции). В случае разбиения области определения на произвольные многоугольники (рис. 8.22, в–д) решение этой задачи отличается крайней сложностью и не найдено до сих пор. Но оно упрощается, если область разбита на четырехугольные или треугольные элементы, и это решение так или иначе основано на идеях теории сплайн-функций и метода конечных элементов.

Кусочно-непрерывные математические модели местности подразделяются на регулярные и нерегулярные.

Из регулярных моделей наиболее часто используются модели на сетке квадратов или прямоугольников (рис. 8.23). Такая модель представляет собой структуру $M = \{X_0, Y_0, H_0, \alpha, S_x, S_y, S_h, m, n, H_{mn}\}$, где X_0, Y_0 – исходные координаты юго-западного угла сетки квадратов; H_0 – значение начальной высоты; α – угол разворота модельной сетки относительно исходной системы координат, S_x, S_y – шаг сетки по X и Y; S_h – шаг квантования высот; m, n – число узлов по X и У; H_{mn} – матрица значений высот в узлах сетки, выраженных в шагах дискретизации. Как правило, сетка прямоугольников параллельна осям координат. При моделировании участков, вытянутых под углом примерно 45 или 135° к оси абсцисс такое представление будет неэффективно, поэтому может вводиться новая система координат, развернутая относительно исходной на угол α.

При использовании нерегулярных моделей область моделирования чаще всего разбивается на непересекающиеся треугольники с вершинами в исходных точках. Такое разбиение называют также плоской триангуляцией, или триангуляционным покрытием, хотя лучше говорить о триангуляционном разбиении. В каждом треугольнике поверхность обычно считается плоскостью,

проходящей через его вершины, в результате чего вся поверхность представляется многогранником с треугольными гранями. Тогда функция (поверхность) всюду непрерывна, но ее производные терпят разрыв на сторонах треугольников. Сегодня это самый распространенный метод построения моделей топографических поверхностей. Представление плоской триангуляции из-за многообразия решений требует особого рассмотрения, что будет сделано далее.

Предлагалось также покрытие области моделирования сеткой непересекающихся произвольных четырехугольников. Сеть треугольников может быть уложена на любой поверхности, но о сетке произвольных четырехугольников этого сказать нельзя, поскольку в ней возможно вырождение четырехугольников в треугольники.

Применение дискретных моделей основано на сформулированной К. Шенноном теореме, в соответствии с которой любая непрерывная на отрезке функция с необходимой точностью может быть представлена конечной последовательностью точек.

Деление моделей на непрерывные и дискретные в определенной мере условно, поскольку даже вещественные числа в ЭВМ являются дискретными величинами. Дискретными далее будем называть модели топографической поверхности, в которых значения высот известны только для конечного множества точек. Если для всего заданного множества точек построить восполняющую функцию, то это будет непрерывная модель. Если на множестве заданных точек построить множество элементов и для каждого элемента определить уникальную восполняющую функцию, то это будет кусочно-непрерывная модель.

Однако, как уже говорилось, математические модели характеризуются в первую очередь своей структурой. Поэтому рассмотрим, какие отношения между точками топографической поверхности воспроизводятся в ее математических моделях.

Единственное отношение, воспроизводимое в аналитической модели топографической поверхности, есть не что иное, как функция h, описывающая топографическую поверхность. Это отношение является функциональным, когда из равенства аргументов следует равенство значений функции. Конечно, точки реальной топографической поверхности не находятся в таком отношении, но если модель признана адекватной, то не остается ничего иного, как признать существование функционального отношения.

Дискретная модель топографической поверхности представляет собой конечное множество точек, и ее важнейшим показателем служит характер распределения точек на поверхности. Выше уже говорилось, что точки моделируемой поверхности идентифицируются своими координатами на поверхности относимости. Не теряя общности в рассуждениях, для удобства поверхность относимости можно считать плоскостью. Тогда по характеру распределения точек на плоскости дискретные модели топографических поверхностей, как и схемы выборки, могут подразделяться на регулярные, полурегулярные и нерегулярные.

Второй важнейшей характеристикой моделей топографических поверхностей служит степень связности. Значение связности таково, что некоторая конструкция из абстрактных или материальных элементов может рассматриваться как целое, как система лишь благодаря существованию тех или иных связей (отношений) между элементами.

Несмотря на признание связности как основополагающей характеристики целого, не существует ее более или менее удовлетворительного общего определения. Нет также единого мнения о том, что может служить мерой связности, хотя интуитивно очевидно, что связность определяется не только количеством связей, но и их качеством и способом их задания.

В «связностном» аспекте математические модели топографических поверхностей ранее, видимо, никогда не рассматривались. Не претендуя на окончательное решение вопроса, в качестве рабочего введем следующее определение. Связностью W математической модели топографической поверхности будем называть среднее число связей, приходящихся на одну точку:

$$W = \frac{1}{n} \sum_{i=1}^{n} w_i, \qquad (8.16.1)$$

где w_i – число связей i-й точки; n – число точек. Определенная таким образом связность слабо зависит от числа исходных точек и представляет собой среднее значение степени вершины графа. В формуле (8.16.1) каждое ребро учитывается дважды, поэтому для вычисления W можно использовать выражение

$$W = \frac{2|R|}{|P|}, \qquad (8.16.2)$$

где $|R|$ – число всех ребер; $|P|$ – число всех вершин графа.

Связность дискретной модели может наглядно демонстрироваться с помощью неориентированного графа. Заданные точки образуют вершины графа, а связи (отношения) – его ребра. Существование отношения между двумя точками обычно означает возможность линейной интерполяции между этими точками или интерполяции, близкой к линейной. Следовательно, между двумя соседними вершинами графа, представляющего дискретную модель, не может быть двух ребер. Отсюда также следует, что граф, построенный на вершинах дискретного множества точек, представляющих однозначную поверхность, является плоским.

Иногда отношения на множестве точек дискретной модели могут не задаваться, то есть модель является вырожденным графом. Это имеет место при нерегулярных схемах выборки, когда точечное множество образовано характерными точками топографической поверхности. Пустое множество отношений между вершинами графа является предельным, вырожденным случаем. Связность таких неупорядоченных дискретных моделей равна нулю.

Между удалением точек друг от друга и возможностью линейной интерполяции существует определенная корреляционная зависимость. Как правило, линейная интерполяция возможна между близлежащими, соседними точками. Указанное свойство позволяет восстанавливать связи между точками с использованием тех или иных формальных критериев. Таким образом, из отношения соседства двух вершин графа с некоторой вероятностью следует существование отношения смежности между этими вершинами.

При стремлении числа точек к бесконечности связность связной нерегулярной модели (плоской триангуляции) стремится к 6

$$\lim_{n \to \infty} W = 6 .$$

Связность регулярных дискретных моделей на сетке квадратов или равносторонних треугольников при увеличении размеров сетки стремится соответственно к 4 или 6. Заметим еще раз, что каждая связь (ребро графа) при этом считается дважды. Наиболее привлекательным свойством упорядоченных моделей является возможность задания связей с помощью описания. Наиболее проста эта связь в моделях на прямоугольной сетке.

В качестве еще одной характеристики структуры связной дискретной модели можно ввести значение *нормированной связности*, которую определим как

$$V = \frac{W}{n-1} \qquad (8.17)$$

или

$$V = \frac{1}{n(n-1)} \sum_{i=1}^{n} w_i , \qquad (8.18)$$

где $\frac{1}{n-1} = K$ является *нормирующим коэффициентом*. Увеличение размеров моделируемой области при постоянной плотности точек и увеличение плотности точек при фиксированных границах области моделирования сопровождается уменьшением значения нормированной связности, что дает возможность разбиения области на отдельные участки и их раздельного моделирования. При соблюдении определенных условий нормированная связность каждого участка будет выше, чем нормированная связность всей области.

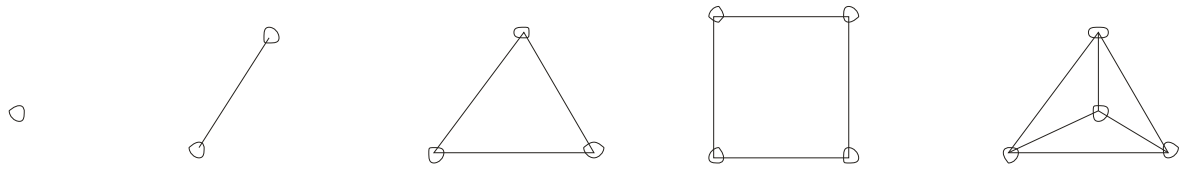
Понятия связности и нормированной связности применимы не только к дискретным математическим моделям, но и к непрерывным. Для вычисления коэффициентов уравнения, описывающего поверхность, составляется система линейных уравнений, число которых равно числу исходных точек. Каждому уравнению ставится в соответствие одна точка, которая связывается в уравнении с остальными точками обычно при помощи некоторой функции координат точки на поверхности. Тогда связность аналитических моделей равна $W = n - 1$, где n – число исходных точек. Отсюда следует, что нормированная связность непрерывных моделей равна 1.

Значение нормированной связности изменяется в диапазоне от 0 (для дискретных несвязанных моделей) до 1 (для непрерывных моделей). Нормированная связность других моделей принимает промежуточные значения, никогда не выходя за указанные пределы.

Своеобразие и соотношение связности и нормированной связности можно проиллюстрировать на следующих простых примерах (рис. 8.24).

1. Дискретная модель состоит из одной точки и $W = 0$. В данном случае не имеет смысла ставить вопрос о значении нормированной связности, поскольку $V = 0/0$.
2. Дискретная связная модель содержит две точки (диполь): $W = 1$ и $V = 1$.
3. Дискретная модель состоит из одного треугольника: $W = 2$ и $V = 1$. (Каждая точка связана с остальными.)
4. Для дискретной модели из одного квадрата: $W = 2$, $V = 2/3$.
5. Для вставки в треугольник (рис. 8.24, *д*): $W = 3$, $V = 1$.

Очевидно, что для любого полного графа значение связности будет $W = n - 1$, а значение нормированной связности $V = 1$.



а) *б*) *в*) *г*) *д*)

Рис. 8.24. Связность примитивов

Таким образом, значение нормированной связности характеризует математическую модель в целом, является ее глобальной характеристикой в отличие от связности, которая суть усредненная локальная характеристика отдельной точки.

Между свойством упорядоченности дискретных моделей и их связностью существуют своеобразные отношения. Упорядоченность характеризует модели со стороны их метрических свойств, а связность описывает их топологию, то есть структуру графа. При сборе информации выбор точек, которые в дальнейшем будут представлять топографическую поверхность, определяется ее внутренней геометрией. При восстановлении поверхности по заданным точкам решается обратная задача – по метрическим свойствам предпринимается попытка воспроизвести ее внутреннюю геометрию.

При рассмотрении совокупности связей между точками дискретной модели выделяются два типа отношений. Отношения первого типа разбивают область моделирования (для простоты будем считать ее односвязной областью) на конечное число подобластей таким образом, что $D = D_1 \bigcup D_2 \bigcup ... \bigcup D_n$ и $D_i \bigcap D_j = \varnothing$, если $i \neq j$. Покрывающие элементы могут быть прямоугольниками, квадратами, правильными или произвольными треугольниками, выпуклыми четырехугольниками или произвольными многоугольниками. Границы этих элементов могут быть криволинейными.

Отношения второго типа служат для задания на поверхности структурных линий, образующих некоторый «каркас», на который «натягивается» поверхность. Отношения обоих типов имеют общие черты: являются двухместными и служат для задания некоторых линий на области определения и/или самой поверхности. Разомкнутая кривая, проходящая через n исходных точек, представляется n – 1 бинарными отношениями, либо одним n-местным отношением, которое получается с помощью склеивания соответствующих бинарных отношений. Замкнутая кривая, содержащая n точек, может быть представлена n бинарными отношениями или одним n-местным отношением. Несмотря на различия в интерпретации отношений первого и второго типов, с формальной точки зрения они эквивалентны.

Отношения между точками регулярных дискретных моделей не представляют особого интереса с точки зрения их построения, так как:

– определение соседних точек является простой задачей;

– связи между точками не являются столь информативными, как это имеет место в нерегулярных моделях.

8.6. Информационные модели топографической поверхности

Математическая модель топографической поверхности – это абстракция, заданная «ни на чем». Рассматривая вопрос о ее реализации, мы приходим к понятию информационной модели топографической поверхности. Далее информационную модель топографической поверхности будем понимать как отображение математической модели топографической поверхности на модель ЭВМ или тройку $D=(M,C,U)$, где M – математическая модель топографической поверхности; C – абстрактное представление ЭВМ; U – отображение $U:M\to C$. Иными словами, информационную модель топографической поверхности далее будем понимать как реализацию ее математической модели на конкретном классе ЭВМ.

Математические модели – идеальные объекты, информационные модели – это эмпирические объекты, совокупности символов, которые могут существовать только как некоторая субстанция. Естественно, что такая субстанция обладает присущими ей свойствами, которых нет у идеального объекта – математической модели.

В терминах обработки данных информационные модели являются определенным составом (содержанием) и структурой данных о некоторой предметной области. Свойства и классификация информационных моделей топографических поверхностей во многом повторяют свойства и классификацию математических моделей поверхностей, но вместе с тем имеют свои особенности.

Подобно математическим, информационные модели подразделяются на непрерывные, или аналитические, и дискретные. При этом классификации математических моделей по типу базового множества соответствует разбиение информационных моделей по содержанию. Содержание аналитических моделей топографических поверхностей составляют параметры (обычно коэффициенты)

некоторого уравнения, описывающего поверхность; содержание дискретных – значения координат и высот земной поверхности.

Структуры непрерывных информационных моделей топографических поверхностей достаточно просты, так как являются повторением структуры математического выражения, на практике всегда представляющего упорядоченную конечную последовательность (кортеж) его параметров.

При использовании нерегулярных дискретных информационных моделей возникают определенные проблемы, связанные с отображением неупорядоченного множества точек двумерного пространства на одномерное, каким является внутренняя или внешняя память ЭВМ. Такие затруднения возникают, например, при поиске вершины, ближайшей к заданной, или нескольких таких вершин. Если множество вершин никак не упорядочено, то для поиска ближайшей придется просматривать все вершины, что оборачивается неэффективным использованием процессорного времени.

Решение подобных проблем, не возникающих в математических моделях, сводится к построению эффективной адресной функции. Размещение образов точек поверхности в ЭВМ можно рассматривать как их упорядочивание. В принципе, точки могут быть упорядочены любым способом, например, произвольным расположением в памяти и нумерацией в возрастающем порядке. Однако, такое «упорядочивание» не дает никаких преимуществ. Поэтому произвольное, но фиксированное преобразование I, ставящее в соответствие образу каждой точки некоторое целое число, будем называть индексированием $I : P \to \{1, 2, ..., |P|\}$, где $|P|$ – мощность базового множества информационной модели. Индексы точек играют роль их указателей – наиболее эффективных адресных функций, и поэтому используются везде, где это только возможно.

Очевидно, что сложность упорядочивания множества образов точек в памяти будет зависеть от того, насколько точки упорядочены на земной поверхности.

Дискретные информационные модели отражают два свойства точек топографической поверхности: плановое положение и высоту. Под упорядочением будем понимать некоторую зависимость между взаимным расположением точек на местности (координатами точки) и взаимным положением их образов в памяти ЭВМ, в общем случае такая зависимость может быть достаточно сложной. Наиболее благоприятные предпосылки для упорядочивания образов точек в памяти ЭВМ возникают при использовании регулярных схем выборки, когда адресная функция представляется аналитически, и положение образа точки определяется с точностью до кванта памяти. Чтобы обеспечить быстрый доступ к образу точки при полурегулярных схемах выборки, последние, по возможности, трансформируются в регулярные. Чтобы ускорить поиск образа точки по ее координатам в моделях с нерегулярной схемой выборки, может осуществляться их упорядочение либо по координатам, либо по квадратам, либо с использованием квадродеревьев.

Наиболее простыми по своей структуре являются неупорядоченные несвязные дискретные модели, представляющие собой множество троек $\{(x_1, y_1, z_1), ..., (x_n, y_n, z_n)\}$, где кортеж (x_i, y_i, z_i) – это координаты точки на поверхности (необязательно прямоугольные) и высота. Порядок следования троек внутри множества безразличен, неупорядоченные модели в виде любых перестановок точек считаются эквивалентными.

Таким образом, метрические свойства схемы выборки, ее регулярность существенно влияют на структуру дискретных информационных моделей топографических поверхностей, и по этому признаку будем подразделять их на упорядоченные (или регулярные), частично-упорядоченные (полурегулярные) и неупорядоченные (нерегулярные).

В топологическом аспекте дискретные модели характеризуются связностью как совокупностью отношений между элементами точечного множества – вершинами графа. В упорядоченных дискретных моделях нет необходимости задавать отношения между вершинами графа перечислением, но это не значит, что они являются обязательно несвязными. Их особенность состоит в том, что информация об отношениях на графе представляется в виде функции соседства, каждому значению аргумента которой соответствует сразу несколько значений функции и, наоборот, значению функции ставится в соответствие несколько значений аргумента. В частности, для сетки прямоугольников или квадратов функцию соседства можно определить как

$$S(P_{ij}) = \{P_{i-1\,j}, P_{ij-1}, P_{i+1\,j}, P_{ij+1}\}.$$

Свойство этой функции в том, что если при значении аргумента P_{ij} выбрать любое из нескольких значений функции, а затем использовать его в качестве аргумента, то среди вновь полученных значений функции будет присутствовать P_{ij}, то есть, если $S(P_{ij}) = P_{kl}$, то $S(P_{kl}) = P_{ij}$. Иначе, если точка P_{ij} является соседней для точки P_{kl}, то и P_{kl} является соседней для P_{ij}. Это означает, что отношение соседства S является симметричным: из $P_i \leftrightarrow P_j$ следует $P_j \leftrightarrow P_i$, где $\leftrightarrow$ – символ отношения соседства.

В регулярных моделях отношение инцидентности для своего представления также не требует перечисления и каких-либо структур данных. Факт инцидентности, например, любого ребра и заданной вершины может определяться аналитически.

Необходимость представления отношений между узлами в явном виде возникает при использовании нерегулярных схем выборки. С формальной точки зрения совокупность бинарных отношений $R \subset \mathrm{P} \times \mathrm{P}$ между вершинами графа (или вершинами и ребрами) представима в виде матрицы смежности или матрицы инциденций. Поэтому первая подходящая машинная структура для отображения множества отношений в памяти ЭВМ – это вектор памяти. Но столь прямолинейный подход к решению задачи на больших объемах данных

может привести к физической невозможности размещения данных из-за ограниченного объема памяти. В лучшем случае память ЭВМ будет использоваться крайне расточительно, так как матрица оказывается сильно разреженной, число ненулевых элементов может составлять менее 1 %.

Матрица смежности и матрица инциденций могут быть использованы для представления отношений на любом графе. Неэффективное использование памяти вычислительного комплекса – неизбежная расплата за универсальность подобного рода. Если возникает проблема оптимизации использования памяти, а на реальных данных она возникает неизбежно, то необходимо учитывать специфику графа. Структура графа предопределена способом интерпретации ребер между его вершинами. Если ребра интерпретируются как границы, разбивающие область моделирования на элементы, то весьма ценной оказывается информация об их форме. Как правило, предполагается разбиение области определения на элементы одного типа: треугольники, выпуклые четырехугольники, произвольные многоугольники. Границы области моделирования при этом совпадают с границами некоторых элементов.

Если покрывающие область моделирования элементы являются многоугольниками, то описание границ элементов с помощью бинарных отношений оказывается неэкономным из-за его информационной избыточности. Этот недостаток устраняется с помощью операции склеивания отношений. Далее можно действовать двумя способами. В первом случае полученное n-местное отношение полностью описывает многоугольник, и каждому многоугольнику ставится в соответствие единственное отношение $G_k = \{\{P_i\}_i^n, \{R_j\}_j^m\}$, а множество всех границ есть объединение $G = \bigcup_{k=1}^{p} G_k$, где p – число всех многоугольников.

Во втором случае отношение описывает цепь между двумя узловыми вершинами. Эти отношения также индексируются. Каждому многоугольнику в соответствие ставится упорядоченный набор отношений, представляющих простые цепи. Иначе, модель границ представляется как $G = \{\{P_i\}_i^n, \{R_j\}_j^m, \{S_k\}_k^l\}$, где $P_i = (x_i, y_i, z_i)$ – образ точки, $R_j = (I_{j1}, ..., I_{jL})$ – индексы вершин, образующих цепь из одной узловой вершины в другую узловую вершину (не содержащую других узловых вершин), $S_k = (J_{k1}, ..., J_{kq})$ – индексы звеньев, в совокупности формирующие границу многоугольника. Реализация этой логической структуры может осуществляться либо на последовательной, либо на связанной памяти.

8.7. Представление плоской триангуляции

Представление информационных моделей топографических поверхностей на нерегулярной сетке треугольников требует особого рассмотрения. Во-первых, эти модели являются наиболее универсальным способом представления поверхностей. Поэтому модели топографической поверхности в

виде нерегулярной сетки треугольников (или плоской триангуляции) являются наиболее распространенными. Во-вторых, задача разработки структуры триангуляции не так проста, как может показаться на первый взгляд.

При представлении триангуляции мы имеем дело с тремя сущностями: вершинами, ребрами и треугольниками – и двумя видами отношений: смежности и инцидентности. Отношение смежности существует между сущностями одного типа: вершина – вершина, ребро – ребро, треугольник – треугольник; отношение инцидентности – между сущностями различного вида (вершина – ребро, вершина – треугольник, ребро – треугольник).

В процессе построения триангуляции и решения различных задач с ее использованием приходится отвечать на следующие вопросы:

какие вершины являются смежными для данной вершины *v*;

какие ребра инцидентны вершине *v*;

какие треугольники инцидентны вершине *v*;

какие вершины инцидентны ребру *e*;

какие ребра являются смежными ребру *e*;

какие треугольники инцидентны ребру *e*;

какие вершины образуют данный треугольник *t* (инцидентны ему);

какие ребра инцидентны треугольнику *t*;

какие треугольники являются смежными для треугольника *t*.

Каждое отношение инцидентности может быть представлено двояко. Так, для каждой вершины могут перечисляться все инцидентные ей ребра, но можно для каждого ребра указывать две инцидентные ему вершины.

Подобные вопросы возникают, например, при перестройке триангуляции. Так, если удаляется некоторая вершина, то необходимо найти и удалить все инцидентные ей ребра и треугольники. Если удаляется ребро, то требуется найти и удалить инцидентные ему треугольники. Для вычисления площади любого треугольника нужно знать его вершины и т. д.

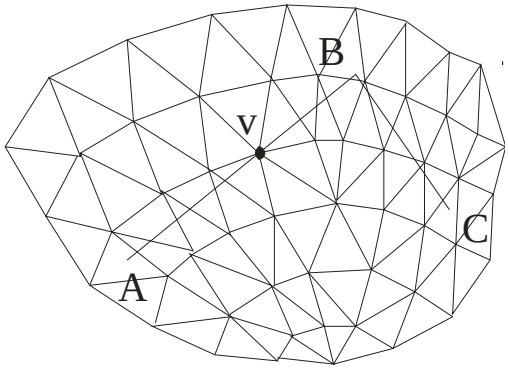


Рис. 8.25. К построению профиля

Кроме того, существует такая задача, как навигация в сети треугольников, возникающая при построении профилей, отслеживании горизонталей по нерегулярной треугольной сетке и т. п. Так, при построении профиля по линии ABC (рис. 8.25) при выходе из треугольника требуется находить следующий смежный треугольник. При прохождении профиля точно через вершину v необходимо просматривать все инцидентные ей треугольники и выбирать из них следующий нужный треугольник.

Представление плоской триангуляции может осуществляться разными способами. Наиболее полно они были описаны, вероятно, в статье Л. де Флориани [35].

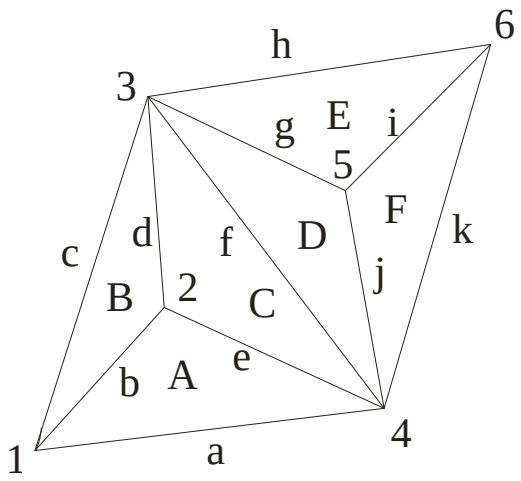


Рис. 8.26. Пример триангуляции

Число треугольников |T| и число ребер |E| в односвязной триангуляции, то есть в триангуляции без дыр, не зависят от способа ее представления и выражаются соответственно формулами

$$|T| = 2(n-1) - m; \qquad (8.19)$$

$$|E| = 3(n-1) - m, \qquad (8.20)$$

где n – число всех вершин триангуляции; m – число граничных вершин.

Для представления триангуляции в памяти машины осуществляется индексация вершин, ребер и треугольников, то есть их отображение на подмножество натуральных чисел $I : V \to \{1, 2, ..., |V|\}$, $J : E \to \{1, 2, ..., |E|\}$, $K : T \to \{1, 2, ..., |T|\}$, где V, E и T – соответственно множество вершин, ребер и треугольников. Далее вершину, ребро или треугольник будем отождествлять с соответствующим индексом i, j или k и обозначать как v_i, e_j и t_k.

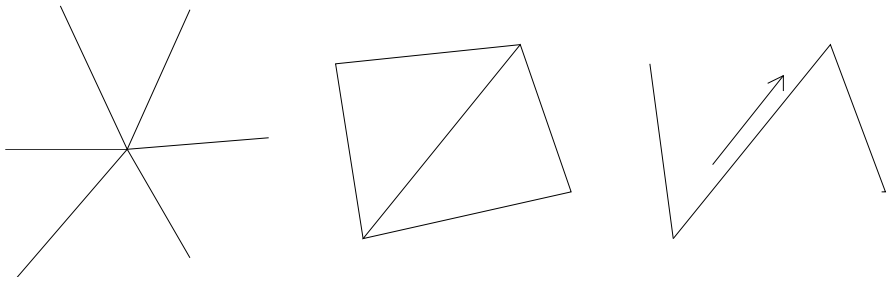
В табл. 8.1 перечислены возможные варианты представления отношений в сети триангуляции. В качестве примера рассматривается триангуляция на рис. 8.26. Вершины триангуляции обозначены арабскими цифрами, ребра – строчными, а треугольники – прописными латинскими буквами. Естественно, в реальных структурах данных им соответствуют индексы (или ссылки, указатели).

В табл. 8.1 число элементов, необходимых для представления отношения, равно 6(n – 1) – 2m или 6(n – 1) – 3m. Так, для хранения отношения смежности ребер необходимо 6(n – 1) – 2m элементов, а для представления отношения смежности треугольников требуется 6(n – 1) – 3m.

Таблица 8.1. Отношения в триангуляции

№ п/п	Отношение	Пример отношения	Структура отношений	Число элементов	
1	Вершина – вершина	2 – 1, 3, 4	$v_i - v_{i1},...,v_{ik}$	2|E|	6(n – 1) – 2m
2	Вершина – ребро	2 – b, d, e	$v_i - e_{i1},...,e_{ik}$	2|E|	6(n – 1) – 2m
3	Вершина – треугольник	2 – A, B, C	$v_i - t_{i1},...,t_{ik}$	3|T|	6(n – 1) – 3m
4	Ребро – вершина	b – 1, 2	$e_i - v_{i1}, v_{i2}$	2|E|	6(n – 1) – 2m
5	Ребро – ребро	b – a, c, d, e	$e_i - e_{i1}, e_{i2}, e_{i3}, e_{i4}$	2|E|	6(n – 1) – 2m
6	Ребро – треугольник	b – A, B	$e_i - t_{i1}, t_{i2}$	2|E|	6(n – 1) – 2m
7	Треугольник – вершина	C – 2, 3, 4	$t_i - v_{i1}, v_{i2}, v_{i3}$	3|T|	6(n – 1) –3m
8	Треугольник – ребро	C – d, e, f	$t_i - e_{i1}, e_{i2}, e_{i3}$	3|T|	6(n–1) – 3m
9	Треугольник – треугольник	C – A, B, D	$t_i - t_{i1}, t_{i2}, t_{i3}$	3|T|	6(n–1) – 3m

Отношение «вершина – вершина» представляется с помощью структуры, называемой барицентрической звездой (рис. 8.27, а). Отношение «ребро – ребро» содержит четыре ребра, смежных данному ребру (рис. 27, б). Но представление данного отношения может быть упрощено, если для каждого ребра хранить не четыре, а только два смежных с ним ребра (рис. 8.27, в). При этом не предполагается, что ребро ориентировано. С первой вершиной ребра связывается первое ребро слева, а со второй его вершиной – первое справа. При смене направления будут указаны эти же ребра.



а) б) в)

Рис. 8.27. Варианты представления отношений

Возможно также использование других структур данных. Например, для каждого треугольника могут указываться номера (индексы) его вершин, ребер и смежных треугольников. Это означает, что отношения 7, 8 и 9 табл. 8.1 объединяются в одно.

8.8. Компактное представление плоской триангуляции

При выборе структур данных необходимо учитывать то обстоятельство, что для решения любых задач на сетке треугольников в представление триангуляции необходимо включать отношения, которые позволяли бы эффективно, без многочисленных переборов переходить от вершин к ребрам, от

вершин к треугольникам и от ребер к треугольникам, а также в обратном направлении.

Кроме памяти, необходимой для хранения отношений инцидентности и смежности в сети триангуляции, при любом способе ее представления требуется пропорциональное 3n пространство для хранения значений координат и высот исходных точек. Таким образом, представление всех данных о триангуляции требует значительных объемов оперативной памяти. Так, если для хранения одного элемента данных любого типа требуется 4 байта, то для представления всей информации о триангуляции в 1 000 точек может потребоваться порядка 10^5 байт и более. Поэтому проблема разработки структур для представления триангуляции продолжает оставаться актуальной.

В формулах числа треугольников и ребер триангуляции (8.19), (8.20) можно заметить, что первое не превышает удвоенного числа вершин, а второе – утроенного числа вершин триангуляции. Поэтому возникает желание найти в хаосе вершин закономерность и каждой точке поставить в соответствие не больше двух инцидентных ей треугольников и не более трех инцидентных ребер таким образом, чтобы их номера были функцией от номера вершины. Тогда эти номера можно не хранить в памяти, а вычислять как функцию от номера вершины.

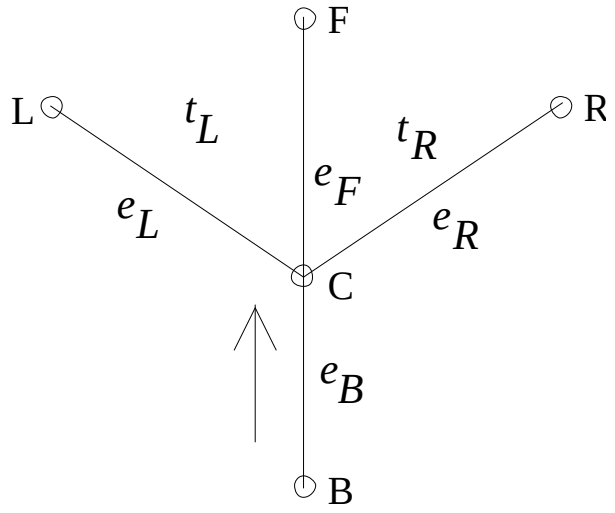


Рис. 8.28. Структура «птичья лапа»

Такую зависимость можно установить, если использовать структуру, представленную на рис. 8.28, где стрелка указывает направление движения, B – задняя, C – текущая, L – левая, R – правая и F – передняя вершины по ходу движения. Последовательность смежных задних и передних вершин назовем траверсом. (В геодезии траверсом иногда называли ход полигонометрии. Мы используем этот термин в силу внешнего сходства изображения последовательности ребер с изображением хода полигонометрии.) На рис. 8.28 последовательность вершин (B, C, F) является участком траверса. А саму структуру, показанную на рис. 8.28, можно назвать птичьей лапой (ПЛ).

Если текущая вершина имеет номер i, то ей можно поставить в соответствие три ребра (e_L , e_F и e_R) и два треугольника (t_L и t_R), примыкающие к ребру e_F . Номера левого и правого треугольников при этом мы можем определить как функции от номера текущей вершины i, например:

$$\left.\begin{aligned} t_L(i) &= 2i-1 \\ t_R(i) &= 2i \end{aligned}\right\}. \qquad (8.21)$$

Подобным же образом можно установить порядок нумерации ребер, например:

$$\left.\begin{aligned} e_L(i) &= 3i-1 \\ e_F(i) &= 3i \\ e_R(i) &= 3i+1 \end{aligned}\right\}. \qquad (8.22)$$

Безусловным преимуществом структуры ПЛ является минимальная избыточность представления данных. В рассмотренных выше структурах для представления триангуляции требуется дублирование одних и те же отношений. Так, например, необходимы указатели как от вершин на ребра, так и обратные указатели от ребер на вершины, если мы хотим осуществлять быстрый поиск в обоих направлениях. Структура ПЛ позволяет хранить не более 4n элементов: указатели на заднюю, левую, правую и следующую вершину траверса. Избыточными в ней являются указатели на заднюю вершину, и их число равно n, то есть по одному на каждую вершину триангуляции. Это число минимально и не так велико по сравнению с другими представлениями.

Рассмотрим применение правил (8.21) и (8.22) на примере представления регулярной триангуляции, изображенной на рис. 8.29, на котором последовательности вершин (2, 5, 4), (6, 7, 1, 9) и (10, 3, 8) являются траверсами. Хотя данная триангуляция является регулярной, ее вершинам номера (индексы) присвоены случайным образом, как это имеет место при нерегулярной триангуляции.

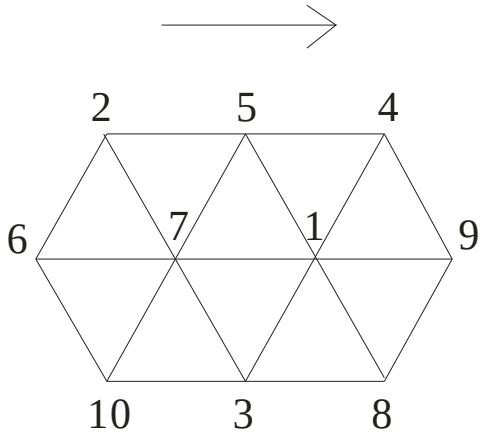


Рис. 8.29. Регулярная сеть

Номера, вычисленные по формуле (8.21) и присвоенные треугольникам, представлены в табл. 8.2. Рассмотрим текущую вершину $v_C = 1$. Для нее $v_B = 7$, $v_L = 4$, $v_R = 8$ и $v_F = 9$. Согласно введенному правилу $t_L = 1$, $t_R = 2$. Для вершины $v_C = 2$ отсутствует левая вершина, признаком чего является $v_L = 0$. Следовательно, отсутствует и левый треугольник, поэтому $t_L = 0$.

Таблица 8.2. Треугольники регулярной триангуляции

Текущая вершина	Задняя вершина	Левая вершина		Правая вершина		Передняя вершина
		v_L	t_L	v_R	t_R	
1	7	4	1	8	2	9
2	0	0	0	7	4	5
3	10	1	5	0	0	8
4	5	0	0	9	0	0
5	2	0	0	1	10	4
6	0	2	11	10	12	7
7	6	5	13	3	14	1
8	3	9	0	0	0	0
9	1	4	0	8	0	0
10	0	7	19	0	0	3

В данном представлении триангуляции наблюдается полная регулярность, за исключением конечных вершин траверсов. Такими вершинами являются 4, 9 и 8. Они не имеют следующих вершин и треугольников, и это свойство может служить признаком конечной вершины траверса. Признаком начальной вершины траверса служит значение предыдущей вершины: $v_B = 0$. Для конечных вершин траверса необходимо указывать либо левую, либо правую вершину. Выберем представление правой вершины. Тогда для вершины 4 правой является $v_R = 9$, а для вершины 9 – $v_R = 8$. Если этого не сделать, то в представлении триангуляции будут отсутствовать ребра (4, 9) и (9, 8).

Но следует еще раз подчеркнуть, что для всех конечных вершин траверсов должны указываться только левые либо только правые вершины. Если указывать и те, и другие, то конечные ребра будут представляться дважды, причем с разными номерами. Перечисление ребер регулярной триангуляции дано в табл. 8.3. В связи с тем, что номера ребер и треугольников являются функцией от номера текущей вершины, при представлении триангуляции с использованием структуры «птичья лапа» нет необходимости указывать эти номера. Тогда регулярная триангуляция, изображенная на рис. 8.29, может быть представлена табл. 8.4.

Таблица 8.3. Ребра регулярной триангуляции

Текущая вершина	Задняя вершина	Левая вершина		Правая вершина		Передняя вершина	
		v_L	e_L	v_R	e_R	v_F	e_F
1	7	4	2	8	4	9	3
2	0	0	5	7	7	5	6
3	10	1	8	0	10	8	9
4	5	0	11	9	13	0	12
5	2	0	14	1	16	4	15
6	0	2	17	10	19	7	18
7	6	5	20	3	22	1	21
8	3	9	23	0	25	0	24
9	1	4	26	8	28	0	27
10	0	7	29	0	31	3	30

Таблица 8.4. Представление регулярной триангуляции

Текущая вершина	Задняя вершина	Левая вершина	Правая вершина	Передняя вершина
1	7	4	8	9
2	0	0	7	5
3	10	1	0	8
4	5	0	0	0
5	2	0	1	4
6	0	2	10	7
7	6	5	3	1
8	3	9	0	0
9	1	4	0	0
10	0	7	0	3

Закономерность регулярной триангуляции проявляется в том, что вершины любого треугольника принадлежат двум траверсам. Два соседних траверса ограничивают ряд или цепь треугольников. Траверсу может принадлежать только одно ребро треугольника. Два другие ребра треугольника соединяют вершины, принадлежащие разным траверсам.

Очевидно, что ориентация плоской триангуляции никаким образом не влияет на структуру данных. Кроме того, регулярность плоской триангуляции можно нарушить с помощью любого топологического преобразования координатной плоскости, что никак не отразится на ее представлении. Примером такой нерегулярной триангуляции служит рис. 8.30, на котором слева изображены горизонтали, а справа – триангуляция, построенная по выбранным на горизонталях точкам. Такое построение триангуляции по горизонталям используется очень часто, поскольку картометрический метод все еще остается распространенным методом создания геоинформационных моделей. Нетрудно видеть, что в этом примере горизонтали играют роль траверсов. Изломанность траверсов никак не влияет на представление триангуляции. Вершины любого треугольника принадлежат двум различным горизонталям, и ни одно ребро треугольника не пересекает какую-либо горизонталь.

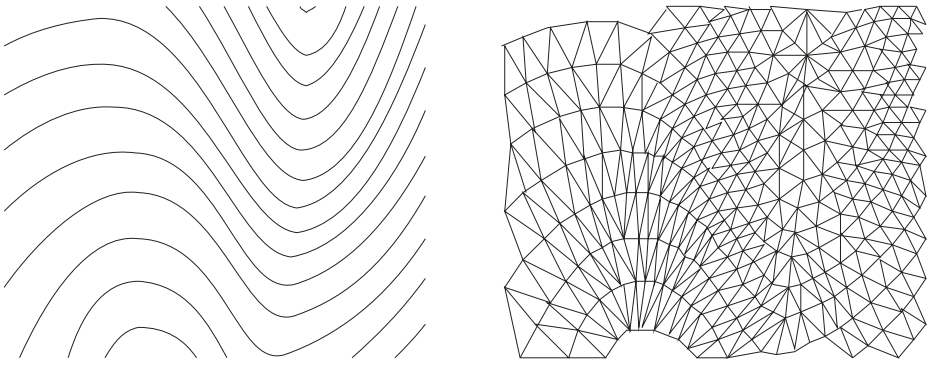
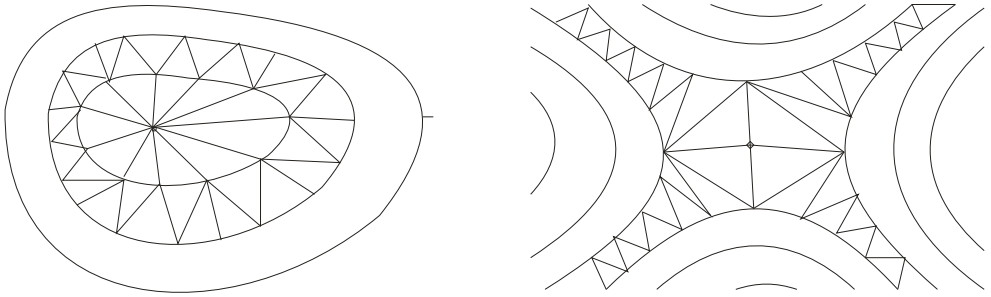


Рис. 8.30. Построение триангуляции по горизонталям

Таким образом, существуют «хорошие» поверхности, которые могут быть описаны с помощью рассматриваемой структуры. Но возникает вопрос, можно ли с ее использованием представить любую триангуляцию? Или, каким

условиям должна отвечать плоская триангуляция, чтобы было возможно использование структуры ПЛ?

Идиллия, подобная представленной на рис. 8.30, нарушается, когда на моделируемой поверхности имеются точки локальных экстремумов (вершины или котловины) и седлообразные участки, а также тогда, когда исходные данные содержат другие характерные точки, не принадлежащие горизонталям. На рис. 8.31 показаны фрагменты триангуляции, построенной на таких участках.



а) *б*)

Рис. 8.31. Нарушения регулярности

При представлении триангуляции с использованием структуры ПЛ возникают некоторые особенности, связанные с представлением вершин локальных экстремумов. Если на рис. 8.31, а обход по траверсу (горизонтали, ближайшей к точке локального экстремума) совершается, например, против часовой стрелки, то для каждой текущей вершины траверса точка локального экстремума будет левой. И здесь никаких осложнений не возникает. Но сама вершина локального экстремума не имеет ссылок ни на одну вершину триангуляции. Она не имеет ни задней, ни передней вершины, поэтому бессмысленно ставить вопрос о левой и правой ее вершинах, поскольку для этой вершины $v_B = 0$, $v_L = 0$, $v_R = 0$ и $v_F = 0$. Но в точке локального экстремума указателю на заднюю вершину можно присвоить значение номера любой смежной с ней вершины. Более сложные проблемы возникают при представлении триангуляции на седлообразных участках (рис. 8.31, б).

В случае регулярной триангуляции структура «птичья лапа» позволяет непосредственно ответить на вопрос об инцидентности любой вершины ребру и треугольнику. Однако в случаях, подобных изображенному на рис. 8.33, нельзя получить такой ответ, обратившись к данным об этой точке. Причина заключается в нарушениях регулярности триангуляции, которые мы будем называть аномалиями регулярности (триангуляции). Они характеризуются тем, что возникают проблемы с представлением некоторых ребер и треугольников.

На рис. 8.32 представлен пример нерегулярной триангуляции, не содержащей аномалий. Следовательно, не любое нарушение регулярности является аномалией. Изображенная на рис. 8.32 триангуляция без каких-либо осложнений может быть представлена с помощью структуры данных ПЛ.

Аномалии регулярности резко ограничивают возможности применения структуры ПЛ. Поэтому нужно определить типы таких аномалий, чтобы модифицировать рассмотренный способ представления триангуляции и сделать его пригодным для любой триангуляции. Одна из аномалий – на концевых вершинах траверсов – уже была рассмотрена выше. Назовем ее аномалией концевых вершин траверсов или аномалией регулярности первого типа.

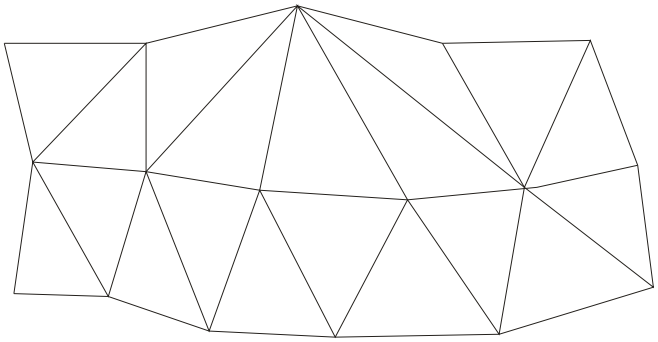
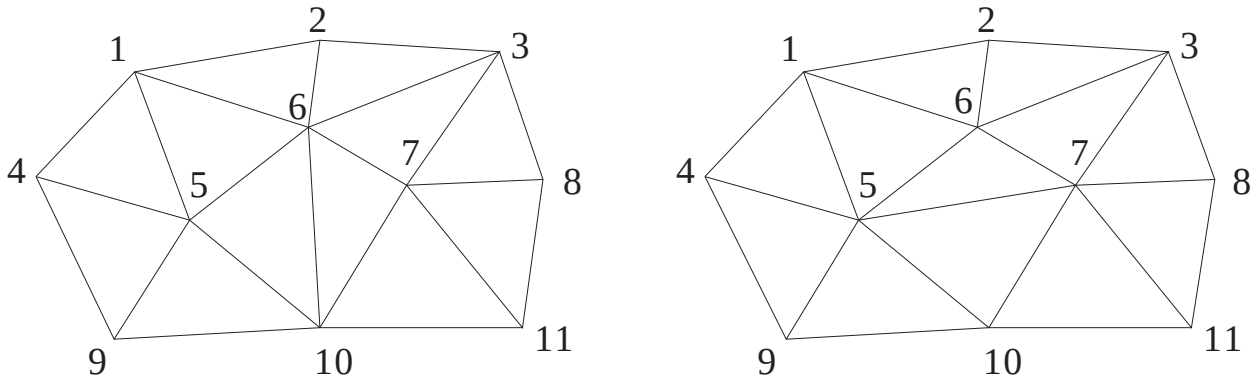


Рис. 8.32. Отсутствие аномалии

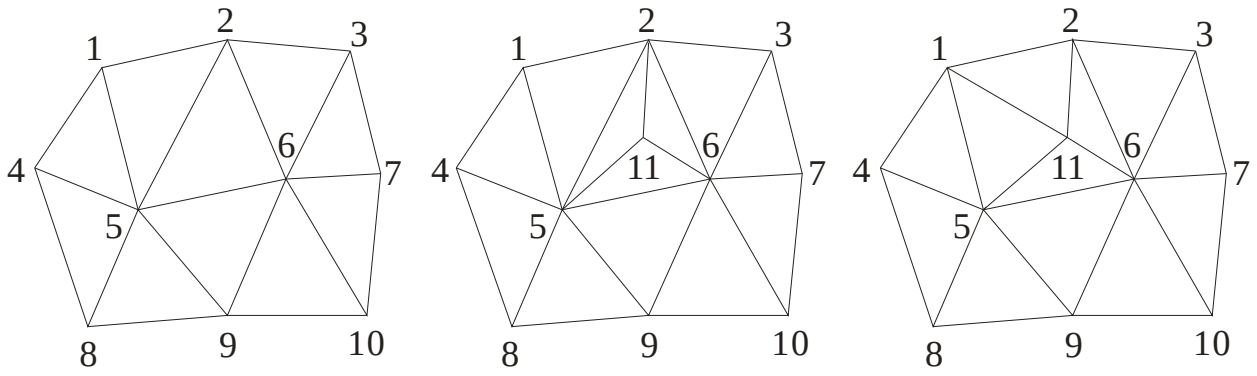
На рис. 8.33 демонстрируется второй случай аномалии регулярности. Слева на нем изображена триангуляция, которая не имеет аномалий и может быть представлена структурой ПЛ. Предположим, что для более точного отображения поверхности необходимо удалить ребро (6, 10) и создать ребро (5, 7). При этом возникает аномалия регулярности: все три вершины треугольника (5, 6, 7) принадлежат одному траверсу (рис. 8.33, *б*).

Еще одна разновидность аномалий регулярности возникает при вставке вершин в существующую триангуляцию. Пусть имелась триангуляция, изображенная на рис. 8.34, *а*, и к ней была добавлена вершина 11. При этом возможна как вставка в существующий треугольник (рис. 8.34, *б*), так и некоторые более сложные перестроения триангуляции, одно из которых представлено на рис. 8.34, *в*, а остальные возможные варианты не показаны.



а) *б*)

Рис. 8.33. Аномалия регулярности типа 2а



а) *б*) *в*)

Рис. 8.34. Аномалия регулярности типа 2b

На рис. 8.33, 8.34 представлены аномалии регулярности одного типа. Если триангуляцию на рис. 8.33 перестроить, заменив траверс (4, 5, 6, 7, 8) траверсом (4, 5, 7, 8), то аномалия регулярности типа 2a сведется к типу 2b. Следовательно, в данном случае мы имеем дело с одним типом аномалий регулярности, который назовем аномалией регулярности типа 2. Его суть заключается в том, что триангуляция дополняется вершиной, лежащей между двумя траверсами и не принадлежащей ни одному из них (рис. 8.35).

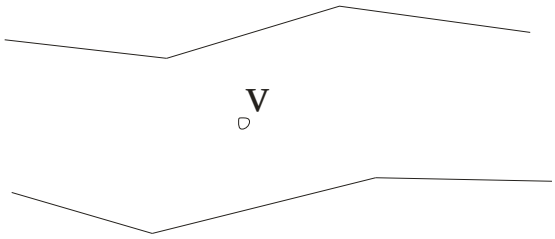


Рис. 8.35. Аномалия типа 2

В общем случае может потребоваться вставка между двумя траверсами не одной, а нескольких следующих друг за другом вершин. Это означает, что на некотором участке между двумя существующими траверсами требуется вставить новый траверс (рис. 8.36). Тогда вставка вершины между траверсами может рассматриваться как вставка траверса, вырожденного в точку. Использование структуры ПЛ для описания триангуляции такого типа не позволяет представить треугольники, отмеченные на рис. 8.36 знаком «+», в начале и конце нового траверса, а также ребро в начале этого траверса, отмеченное знаком «–»; они «исчезают». Все остальные треугольники и ребра представляются без каких-либо осложнений.

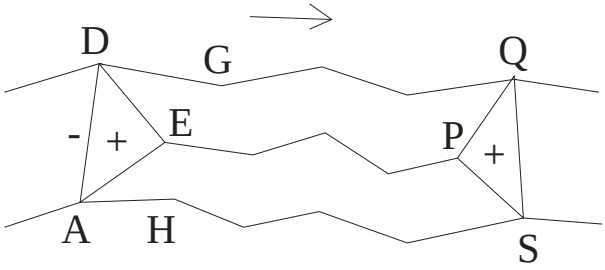


Рис. 8.36. Вставка траверса

Рассмотрим вначале конечную вершину вставляемого траверса. Выше мы говорили, что если указатель на переднюю вершину траверса равен нулю, то это является признаком конечной вершины траверса. Можно принять соглашение, что для представления ребер *PQ* и *PS* в конце траверса (см. рис. 8.36) используется то же правило, что было сформулировано выше. Тогда для

текущей вершины P будет иметь место $v_L(P)=Q$, $v_R(P)=S$, а треугольник является единственным и его номер можно принять равным $2P$.

С начальной вершиной траверса дела обстоят значительно хуже. Для текущей вершины E существуют левая и правая вершины (на рис. 8.36 соответственно G и H). Все связанные с ней указатели на смежные вершины будут иметь фактические значения. Но структура ПЛ не позволяет представить ребро AD.

Заметим, что с любой текущей вершиной связаны два треугольника и три ребра. Но в конечной вершине траверса используются номера только двух ребер и одного треугольника. Таким образом, в концевой вершине траверса остаются свободными недостающие в начальной вершине нового траверса один указатель на треугольник и один указатель на ребро.

Возникшую ситуацию с аномалиями можно рассматривать несколько иначе и интерпретировать как проблему не вставки траверсов, а бифуркации (разделения) и слияния траверсов. На рис. 8.37 слева показана бифуркация, справа – слияние двух траверсов. Как нетрудно видеть, бифуркация и слияние траверсов являются зеркальным отображением друг друга. Можно также сказать, что тип этих аномалий зависит от направления движения, и при его смене бифуркация превращается в слияние и наоборот. На том же рисунке внизу представлен остов этой триангуляции в виде ее траверсов. Вершину, в которой происходит разделение траверса на два траверса, будем называть вершиной бифуркации (траверсов). Вершину, в которой происходит слияние траверсов, назовем вершиной слияния.

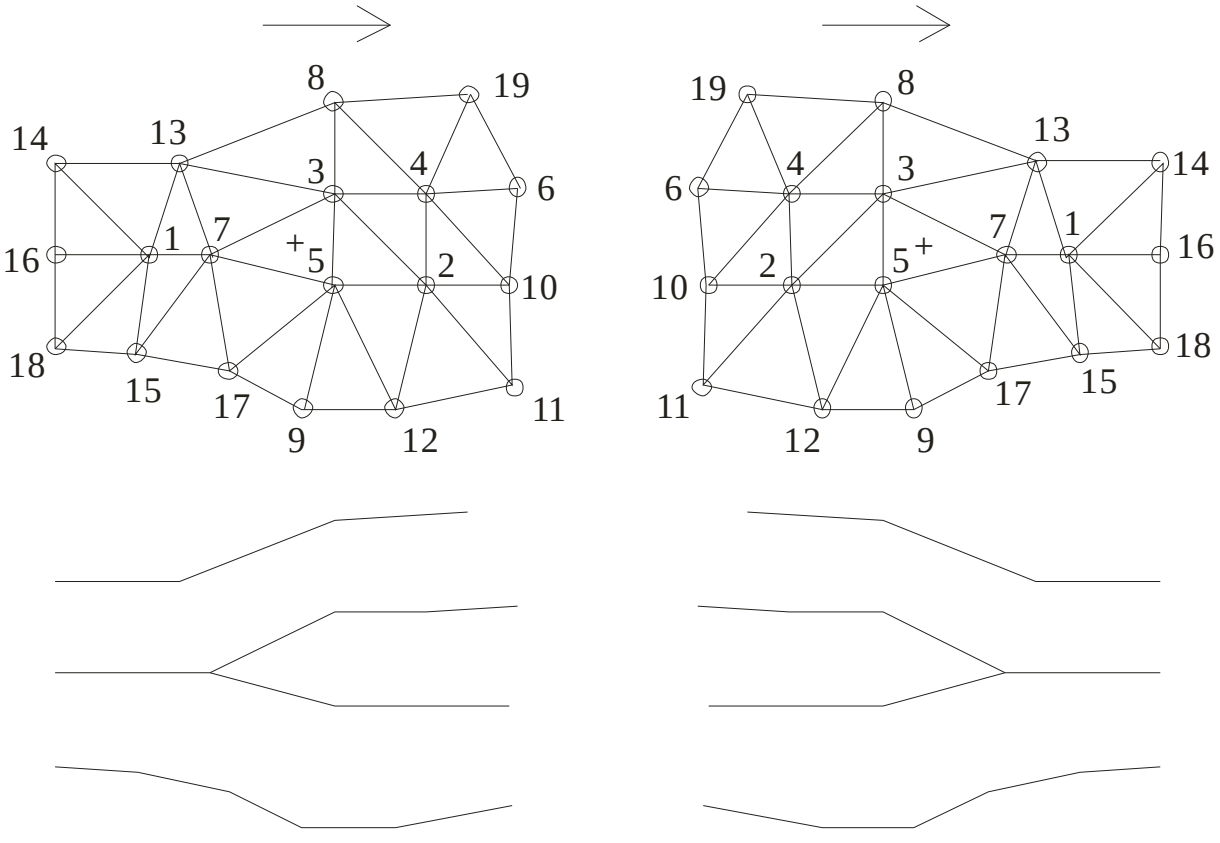
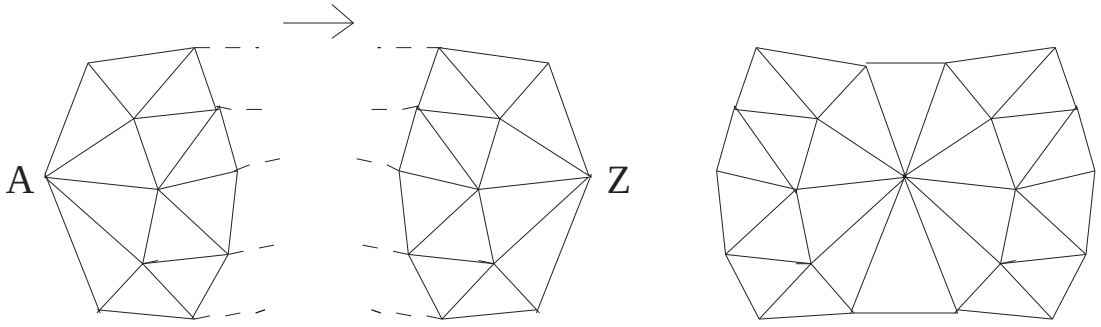


Рис. 8.37. Бифуркация и слияние траверсов

Возможны более сложные случаи бифуркации и слияния траверсов. На рис. 8.38 слева представлена вершина *A*, из которой исходит несколько траверсов (направление движения указано стрелкой). Будем называть подобные точки *вершинами n-бифуркации*, или *множественной бифуркации*. На рис. 8.38, *б* показана *вершина n-слияния*, или *множественного слияния*. Подобные ситуации могут возникать чаще всего в начале и конце траверсов. На рис. 8.38, *в* показана внутренняя вершина триангуляции, одновременно являющаяся вершиной слияния и бифуркации.



а) *б*) *в*)

Рис. 8.38. *n*-бифуркация и *n*-слияние

В триангуляции на рис. 8.37 особенности возникают при представлении треугольников, отмеченных знаком «+». В данном случае вновь требуется принять однозначное решение о порядке присваивания номеров ребрам и треугольникам. Можно, например, ввести следующее уточняющее правило. В вершине бифуркации в качестве передней вершины указывается вершина, принадлежащая левому траверсу. Но что делать с вершиной, находящейся правее, или с несколькими вершинами в случае *n*-бифуркации?

Чтобы обойти проблемы, связанные с бифуркациями и слияниями, можно попытаться перестроить триангуляцию так, как показано на рис. 8.39, на котором по тому же множеству вершин и ребер, что и на рис. 8.38, построена триангуляция, не содержащая вершин бифуркации и слияния. Однако такого рода переструктуризация затронет всю триангуляцию, что потребует больших затрат процессорного времени в больших триангуляциях.

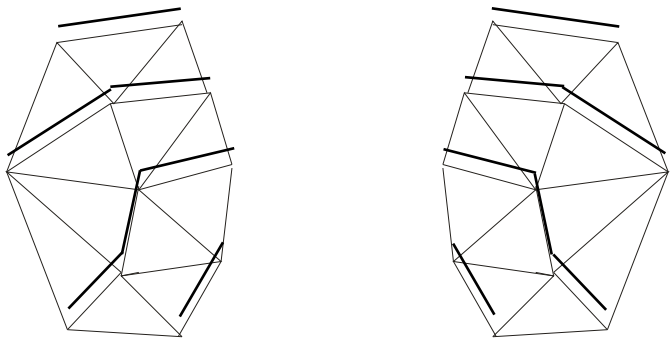


Рис. 8.39. Отсутствие аномалий

Если в триангуляции отсутствуют бифуркации и слияния, то такая триангуляция может быть представлена структурой ПЛ. Но если в триангуляции появляются вершины бифуркации, то она уже не может быть представлена без переструктуризации. Рассмотренные выше случаи вставки вершин между траверсами влекут за собой бифуркацию и слияние траверсов. Таким образом, проблема корректного представления

триангуляции сводится к задаче представления бифуркаций и слияний траверсов.

В точках слияния (см. рис. 8.38) в каждом треугольнике две стороны принадлежат разным траверсам. И такой треугольник представляется дважды, так как он примыкает и к левому, и к правому траверсу. Поэтому при вычислении номера треугольника будет получено два значения. Первое из них будет функцией точки на левом траверсе, а второе – функцией вершины на правом траверсе. Чтобы решить проблему неоднозначности, можно ввести таблицу синонимов треугольников. Тогда номера треугольников будут просто разными именами одного и того же объекта – треугольника.

Можно условиться считать, что треугольник примыкает только к левому траверсу. Тогда каждый раз при вычислении номера треугольника необходимо проверять, не примыкает ли он к левому траверсу. И если ответ на этот вопрос положительный, то номер этого треугольника следует вычислять как функцию от номера вершины на левом траверсе.

В конечном итоге проблема состоит в том, что из точек бифуркации исходит ребер больше, чем их может быть представлено с помощью структуры ПЛ. Таким образом, дело сводится к представлению множественности значений номеров ребер и треугольников, инцидентных вершинам бифуркации, каждая из которых имеет только один номер. Следовательно, вершине бифуркации следует поставить в соответствие не один номер, а несколько номеров, являющихся ее синонимами или псевдонимами. Число таких синонимов в вершинах бифуркации и только в них может увеличиваться по мере необходимости.

Если вершине присвоено два имени (номера, индекса), то в таблице синонимы ссылаются друг на друга. Если у некоторой вершины синонимов более двух, то они образуют кольцевую структуру (рис. 8.40). При этом синонимы играют роль указателей списка. Новому синониму для некоторой вершины присваивается номер на 1 больше, чем число номеров вершин, включая синонимы. После чего число номеров вершин триангуляции (а не число вершин триангуляции) увеличивается на 1. Синоним с наименьшим значением можно считать «действительным» именем вершины. По этому номеру должны выбираться значения координат и высот из перечня исходных точек. Создание синонимов для вершин влечет за собой увеличение накладных расходов на представление триангуляции, но они не слишком большие. Важно, что общий объем памяти, используемой для представления триангуляции, при этом уменьшается.

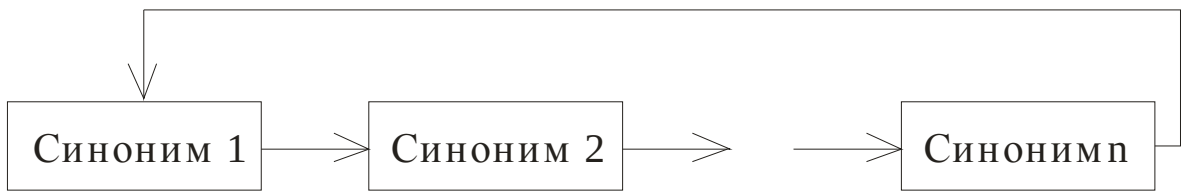


Рис. 8.40. Представление синонимов вершин

Не отказываясь от идеи синонимов, можно принять несколько другое решение и не вводить отдельную таблицу синонимов, а расширить структуру ПЛ, добавив к ней поле «синоним». Тогда новая структура P данных об отдельной вершине триангуляции, которую назовем компактным представлением триангуляции (КПТ), будет иметь вид $P = (C, S, B, L, R, F)$, где C – номер (индекс) текущей вершины; S – синоним текущей вершины; B – задняя вершина траверса для текущей вершины; L – левая вершина (на левом траверсе); R – правая вершина; F – передняя вершина. Если значение S = 0, то это служит указанием на отсутствие синонимов для данной вершины.

Компактное представление триангуляции дает возможность отказаться от создания таблиц синонимов треугольников и ребер, поскольку синонимы и тех, и других могут вычисляться как функции от синонимов вершин. Представление всей триангуляции в таком случае содержит перечень (каталог) координат и высот вершин триангуляции и совокупность всех структур P, включая синонимы.

Представление триангуляции лучше всего начинать с некоторой «центральной» вершины, вокруг которой создается первый траверс, затем все последующие в порядке их удаления от центральной вершины. Движение по всем траверсам осуществляется только в одном направлении: либо по часовой стрелке, либо – против.

Рассмотренное компактное представление не было реализовано, и его реализация является достаточно сложной задачей. Однако, любую программу пишут один раз, а используется она многократно, что и дает экономический эффект.

8.9. Сравнительный анализ информационных моделей

Основными критериями для оценки информационных моделей топографических поверхностей служат адекватность, точность, плотность, объем памяти ЭВМ, процессорное время, стоимость получения, стоимость хранения и удобство использования, называемое иногда открытостью моделей.

Информационные модели могут сравниваться друг с другом, прежде всего, как модели. И тогда можно говорить о степени их адекватности. В конечном итоге информационные модели любой предметной области представляют собой совокупность данных. И тогда разные информационные модели могут сравниваться между собой как данные.

Адекватность любой информационной модели в общем случае определяется такими факторами, как старение, ошибочность и неполнота информации [19]. Рассматриваемые здесь информационные модели топографической поверхности являются статическими. Уже само их назначение заключается в отображении состояния топографической поверхности на определенный момент времени. И в этом смысле статические модели земной поверхности являются своего рода ее «моментальными» снимками. Поэтому старение информации никак не объясняется свойствами модели, вследствие чего этот фактор рассматриваться не будет. Неполнота информации также не зависит от свойств информационной модели, а определяется особенностями

метода сбора информации в конкретных условиях (аппаратура, методика сбора, квалификация и добросовестность исполнителей и т. п.). Следовательно, и этот фактор можно исключить из дальнейшего рассмотрения.

Тогда с позиций нашего исследования адекватность и точность являются синонимами, и адекватность цифровых моделей топографических поверхностей может формулироваться исключительно в терминах точности. Трудно представить, что модель, обеспечивающая вычисление высоты в любой точке с точностью, допустим, 0,01 м, может быть признана неадекватной (по крайней мере, на момент ее создания).

Точность является функцией от плотности, информативности (презентабельности), точности измерения координат и высот исходных точек, метода создания модели, плотности точек модели (для дискретных моделей) либо числа членов в ряде (для аналитических моделей), способа восстановления высот по дискретной модели. Среди этих факторов лишь плотность точек дискретной модели и длина ряда являются внутренними свойствами модели, остальные – внешние факторы. Чтобы иметь возможность сравнивать непрерывные и дискретные модели, нужно число узлов и число членов ряда заменить их общим эквивалентом – объемом памяти ЭВМ.

Если из рассмотрения исключить этап сбора данных, то стоимость создания информационной модели определяется в основном затратами машинного времени, которые в свою очередь зависят от вычислительной сложности применяемого метода моделирования (о конкретных методах моделирования см. ниже). Следовательно, в систему критериев целесообразно включить два показателя: точность и объем памяти ЭВМ.

Теоретически оба типа информационных моделей, то есть, и дискретные, и непрерывные, позволяют достичь любую наперед заданную точность. Однако при реализации математических моделей мы сталкиваемся с конечной точностью представления арифметических данных в ЭВМ.

Увеличение числа параметров в аналитическом выражении обычно характеризуется неприятными последствиями. В частности, при использовании полиномов с возрастанием степени появляются нежелательные осцилляции, и происходит увеличение абсолютных значений их коэффициентов [15]. Но так как точность представления чисел в ЭВМ ограничена, то при увеличении степени полиномов может быть достигнут предел, после которого точность вычисленного значения станет убывать. Кроме того, снижение точности будет происходить из-за увеличения количества операций над приближенными числами, что справедливо для любого аналитического выражения.

Если для представления поверхности требуется хранить только значения коэффициентов уравнения, то такие модели являются самыми компактными: нужно не более n квантов памяти, где n – число исходных точек. Но, как правило, дополнительно требуется запоминание образов исходных точек, так как в уравнении поверхности обычно фигурируют разности координат текущей и исходных точек. Тогда объем используемой памяти возрастает до 4 n квантов. Именно это значение характеризует реальные потребности аналитических моделей в оперативной памяти.

При благоприятном стечении обстоятельств (спокойный характер рельефа и/или малые размеры моделируемой поверхности) число исходных точек составляет несколько десятков или сотен. В подавляющем большинстве случаев оно составляет несколько тысяч, а при моделировании сложных участков местности превышает 104. Поэтому попытки использования непрерывных моделей в реальных условиях обречены на провал, как не обеспечивающие необходимую точность.

Если говорить о потребительских свойствах, удобстве использования разнотипных информационных моделей топографических поверхностей, то в первую очередь следует различать регулярные и нерегулярные модели. Безусловно, что регулярные модели намного превосходят нерегулярные по удобству их использования. Все задачи на регулярных моделях решаются наиболее просто.

Исключение составляет задача картографического отображения топографической поверхности. Об информационных моделях известно, что любая из них характеризуется определенной точностью. Поэтому некоторые погрешности моделей игнорируются при решении инженерных задач. Однако картографическое изображение топографической поверхности, получаемое по информационной модели, должно быть безукоризненным. Оценивая качество изображения рельефа, профессиональный картограф понимает, что он представлен с некоторой точностью, но не может принять внутренней несогласованности элементов изображения друг с другом. Поэтому обычно выполняется так называемая «укладка» горизонталей, даже если она нарушает положение отдельных горизонталей. Таким образом, в отличие от топографов и аэрофотогеодезистов, картографы склонны наделять критерий эстетичности более высоким приоритетом по сравнению с критерием точности. И с этой точки зрения наиболее предпочтительными оказываются нерегулярные модели топографических поверхностей на сетке треугольников.

Причина такого положения в том, что вершинами сетки треугольников, как правило, являются исходные точки, представляющие собой характерные точки поверхности или принадлежащие структурным линиям на ней. Следовательно, сетка треугольников в принципе может быть вписана в моделируемую поверхность. Но регулярные модели таким свойством не обладают. Расположение узлов регулярной модели, например, сетки квадратов, никак не связано со структурными линиями и точками топографической поверхности. Поэтому характерные точки и линии поверхности на картографическом изображении, полученном по регулярной модели, могут смещаться относительно своего истинного положения, в результате чего все изображение будет «корявым». Чтобы получить корректное изображение моделируемой поверхности, при его построении вместе с регулярной моделью необходимо использовать данные о структурных линиях и точках на ней, что усложняет задачу.

Таким образом, право на существование имеют и регулярные, и нерегулярные модели. Регулярные модели находят большее применение при решении различных задач на поверхности, а нерегулярные – при ее

картографическом отображении. Поэтому условно первые можно считать инженерными, а вторые – картографическими. Выбор регулярных или нерегулярных моделей зависит от типа решаемых задач. В конечном итоге он сводится к ответу на вопрос: нужно или нет точно передавать формы рельефа или достаточно только значения высоты (например, для определения поправок за рельеф в значение силы тяжести при гравиметрических съемках).

Сравнивая непрерывные и дискретные модели по частоте их использования, можно прийти к выводу, что последние в «чистом» виде используются крайне редко. Регулярные дискретные модели применяются, например, для вычисления поправок за рельеф при определении значения силы тяжести. В подавляющем большинстве случаев регулярные модели используются для получения кусочно-непрерывных моделей, когда осуществляется их восполнение теми или иными аналитическими методами. Множество точек дискретной модели при этом играет роль «каркаса», на который натягивается поверхность. Наиболее часто используемые дискретные модели – это множество точек, полученных в результате съемки и служащих данными о топографической поверхности или ее первичной моделью.

Если сравнивать непрерывные и дискретные модели по точности, то теоретически и те, и другие могут обеспечить любую практически необходимую точность. Однако, если рассматривать кусочно-непрерывные и дискретные модели на одном и том же множестве точек, то первые отличаются более высокой точностью.

Сопоставляя непрерывные и кусочно-непрерывные модели по области их применения, надо сказать, что непрерывные модели были связаны с самыми первыми попытками моделирования топографических поверхностей и в настоящее время почти полностью вышли из употребления. Но математические методы построения непрерывных поверхностей служат теоретической основой для создания кусочно-непрерывных поверхностей.

Сравнение различных информационных моделей топографических поверхностей по сложности или стоимости их создания мы здесь делать не будем, поскольку эти параметры не относятся к самим моделям. Две модели с одной и той же структурой могут быть получены разными методами, логическая сложность и точность которых, а также вычислительные затраты при их применении могут сильно различаться. На этих характеристиках мы остановимся при рассмотрении методов моделирования топографических поверхностей.

Рассмотрим теперь информационные модели топографических поверхностей в даталогическом аспекте, то есть, как данные. В современных ЭВМ представление данных зависит от их типа. Поэтому сопоставление различных информационных моделей топографических поверхностей означает сравнение данных по способу их представления в ЭВМ.

Параметры уравнения поверхности в аналитических моделях и элементы базового множества в дискретных моделях (значения координат и высот) традиционно задаются вещественными числами. Выбор способа представления параметров уравнения сводится к выбору между вещественными числами

обычной точности (4 байта) или удвоенной точности (8 байт). Целые числа обычно используются в связных моделях как указатели для представления разного рода отношений между элементами базового множества.

В дискретных моделях целые числа могут быть использованы также для представления координат и высот. Для этого необходимо выполнить дискретизацию области определения и квантование области значений функции – высот (так называемое «квантование по уровню»).

Пусть координаты исходных точек даны в системе координат OXY. Дискретизация области определения заключается во введении системы целочисленных координат oxy и выполняется следующим образом:

– определяются минимальные и максимальные значения координат исходных точек $(X_{\min}, X_{\max}, Y_{\min}, Y_{\max})$;

– выбирается шаг *s* дискретизации значений координат, обеспечивающий необходимую точность их представления; возможен выбор значений шага дискретизации s_x и s_y по каждой оси координат такой, что $s_x \neq s_y$;

– начало координат переносится в точку с координатами $(X_{\min} - \frac{s_x}{2}, Y_{\min} - \frac{s_y}{2})$;

– вычисляются целочисленные координаты каждой точки, выраженные в дискретных шагах:

$$x_i = entier(\frac{X_i - X_{\min}}{s_x} + 0{,}5);$$

$$y_i = entier(\frac{Y_i - Y_{\min}}{s_y} + 0{,}5),$$

где entier – функция, вычисляющая целую часть числа, и где 0,5 прибавляется с целью округления до ближайшего целого числа. Аналогичным образом выполняется квантование высот

$$z_i = entier(\frac{Z_i - Z_{\min}}{s_z} + 0{,}5).$$

Дискретизацию области моделирования и квантование высот по уровню не следует воспринимать как экстравагантность. Хотя область моделирования и область значений высот непрерывны, точность любых измерений ограничена практическими потребностями, а также аппаратурными и прочими ошибками. Поэтому на практике значения координат и значения высот всегда округляются до некоторых единиц измерения и представляются с некоторым фиксированным числом знаков. Так, при крупномасштабных топографических съемках значения координат и высот представляются обычно с точностью до 0,01 м, что может рассматриваться как шаг дискретизации и шаг квантования. Таким образом, в геодезии и топографии дискретизация и квантование неявно всегда использовались и используются.

Расчеты показывают, что при использовании целочисленного способа представления координат и высот в дискретных моделях двух байтов вполне достаточно, чтобы обеспечить необходимую практическую точность при моделировании топографических поверхностей, включая автоматизированное составление топографических карт и планов.

Несмотря на очевидную возможность представления дискретных цифровых моделей топографических поверхностей в целочисленном виде, прецеденты практической реализации этой идеи довольно редки. Так, система моделирования топографических поверхностей на целочисленной арифметике для мини-ЭВМ в 1970-х гг. была разработана Топографической службой армии США. Во втором случае метод билинейных сплайнов на подпространстве (см. далее), но не вся система моделирования топографических поверхностей, был реализован на целочисленной арифметике с использованием двухбайтового представления для ЕС ЭВМ в 1980-х гг. при разработке автоматизированной системы картографирования (АСК-1) в НИИПГ.

Заканчивая анализ свойств информационных моделей топографических поверхностей, следует остановиться еще на одной их характеристике, которой математические модели не обладают. Информационная модель топографической поверхности является отображающей системой и, как таковая, может характеризоваться разрешающей способностью, чувствительностью и полосой пропускания. Разрешающая способность и чувствительность информационной модели характеризуют принципиальную возможность воспроизведения локальных особенностей топографической поверхности, а полоса пропускания – диапазон представления координат и высот (или глубин). Разрешающая способность связана со способностью информационной модели правильно отображать горизонтальное простирание отдельных форм топографической поверхности, а чувствительность – со способностью передавать изменения высот, точностью представления их значений.

Пояснить понятие разрешающей способности можно с помощью рис. 8.41, на котором сетка квадратов представляет регулярную дискретную информационную модель топографической поверхности, а замкнутые кривые можно рассматривать как горизонтали, изображающие, например, формы микрорельефа. Из рисунка следует, что микроформы *a* и *b* при заданном шаге сетки квадратов (сплошные линии) не найдут своего отражения в модели. Поскольку размеры этих микроформ хотя бы в одном направлении несколько меньше стороны квадрата, они как бы «провалятся» сквозь сетку квадратов. Микроформы *c* и *d* будут отражены в модели, хотя и в искаженном виде. Из рис. 8.41 следует очевидный способ повышения

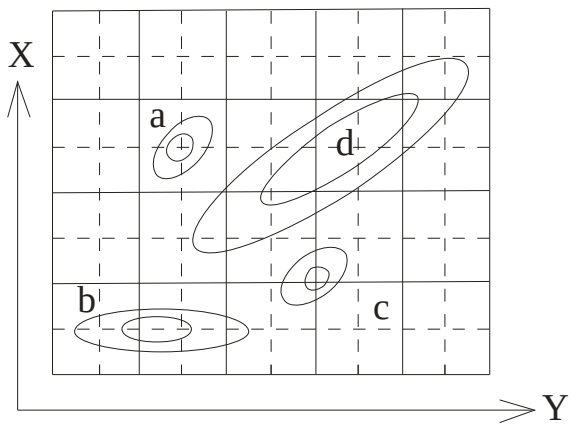


Рис. 8.41. Разрешающая способность

адекватности дискретных регулярных моделей – увеличение их разрешающей способности, то есть уменьшение размеров их элементов. Так, при уменьшении шага сетки в два раза (штриховые линии на рис. 8.41) проявятся микроформы *a* и *b*.

Следовательно, разрешающую способность R регулярной дискретной модели в виде сетки квадратов можно считать равной длине сторон ее квадратов, то есть $R = s$. Делать разрешение различным по осям абсцисс и ординат допустимо, но обычно для этого нет оснований.

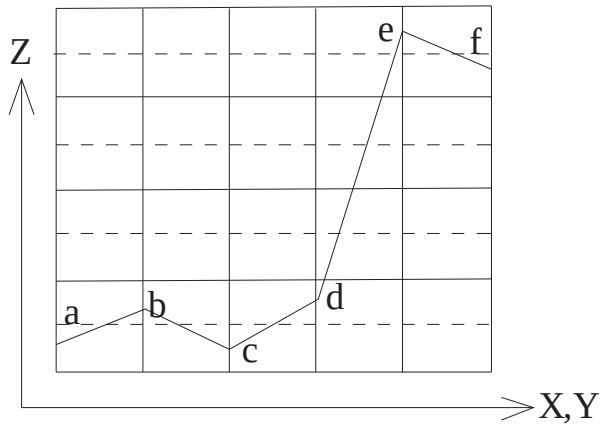


Рис. 8.42. Чувствительность

При низкой разрешающей способности некоторый участок местности может восприниматься как единая форма или совокупность нескольких крупных форм. При увеличении ее значения происходит проявление более мелких деталей, фрагментов.

При фиксированной разрешающей способности с увеличением чувствительности также может происходить детализация общей картины, но только до определенного предела, обусловленного разрешающей способностью. На рис. 8.42 при выбранном значении чувствительности точки a, b, c, d будут иметь одно значение высоты, а точки e и f – другое значение. При увеличении чувствительности, то есть уменьшении шага квантования высот в два раза, происходит проявление более мелких деталей рельефа. Так, на рис. 8.42 значения высот, например, точек a и b будут уже различаться, точка b при этом будет восприниматься как точка локального экстремума.

Таким образом, информационные модели характеризуются ограничениями реального мира – моделирующей системы. В математических моделях, являющихся идеальными объектами, число параметров и точность представления вещественных чисел могут быть сколь угодно большими и в этом смысле математические модели обладают неограниченными возможностями. Точность или адекватность отображения топографической поверхности дискретной моделью при прочих равных условиях (сложности поверхности, плотности исходных точек, их презентабельности, точности определения координат и высот) является функцией от разрешающей способности и чувствительности модели. Разрешающая способность и чувствительность дискретной модели топографической поверхности должны быть определенным образом сбалансированы, согласованы друг с другом.

Характеризуя полосу пропускания, будем различать полосу пропускания в плане, или горизонтальную полосу пропускания, и полосу пропускания высот, или вертикальную полосу пропускания. Горизонтальная полоса пропускания, или полоса пропускания координат g, определяет размеры максимального

участка топографической поверхности, который может быть представлен единой моделью при фиксированном разрешении

$$g_x = X_{\max} - X_{\min};$$

$$g_y = Y_{\max} - Y_{\min}.$$

С другой стороны, горизонтальная полоса пропускания равна

$$g = (2^n - 1)s,$$

где n – число двоичных разрядов в представлении целых чисел; s – шаг дискретизации координат. Отсюда можно получить значение шага дискретизации координат при заданных размерах области моделирования и известной длине целых чисел для представления координат:

$$s = \frac{\max(g_x, g_y)}{2^n - 1}.$$

Вертикальная полоса пропускания v, или полоса пропускания высот, характеризует максимальный диапазон высот, который может быть отображен при заданной чувствительности модели

$$v = Z_{\max} - Z_{\min};$$

$$v = (2^m - 1)s_z,$$

где m – разрядность двоичных целых чисел, используемых для представления высот. Если разрядность используемых чисел известна и известен диапазон изменения высот, то значение шага квантования высот можно получить по формуле

$$s_z = \frac{(Z_{\max} - Z_{\min})}{2^m - 1}.$$

Таким образом, разрешающая способность и чувствительность связаны не только друг с другом, но и с полосой пропускания. При фиксированной длине кванта памяти для хранения координат и высот точек увеличение разрешающей способности (чувствительности) ведет к уменьшению горизонтальной (вертикальной) полосы пропускания, что не всегда может оказаться желательным. В дискретных моделях топографических поверхностей, реализованных с использованием четырехбайтового представления целых чисел, проблемы отношения чувствительности и полосы пропускания не существует. Однако она сразу возникает, как только для повышения эффективности путем минимизации используемых ресурсов ЭВМ осуществляется переход к целочисленной арифметике и двухбайтовому представлению.

В разных типах информационных моделей топографических поверхностей понятия разрешающей способности и чувствительности различаются своим содержанием. В дискретных моделях две точки различимы до тех пор, пока различаются их образы в памяти ЭВМ. Поэтому их разрешающая способность связана с точностью представления координат, которая при использовании

вещественной арифметики интерпретируется как разрядность кванта памяти либо как шаг дискретизации координат. Чувствительность дискретных моделей может быть проинтерпретирована аналогичным образом, а также может быть выражена в терминах высоты сечения рельефа горизонталями.

Разрешающая способность непрерывных моделей также выражается точностью представления координат, а чувствительность является некоторой функцией от точности представления аргументов и точности представления параметров аналитического выражения, описывающего поверхность. Чувствительность непрерывной модели в пределах одного участка моделирования является переменной величиной, тогда как чувствительность дискретных моделей может быть постоянной.

Общая закономерность заключается в том, что увеличение чувствительности, разрешающей способности и полосы пропускания ведут к повышению адекватности (точности) создаваемых моделей при одновременном возрастании необходимого объема памяти. Проблема заключается не в моделировании топографических поверхностей, а в их эффективном моделировании, в использовании минимума необходимых вычислительных ресурсов для достижения заданной точности.

Возможно, что с точки зрения оптимизации процесса получения топографических карт эффективность регулярных моделей может ставиться под сомнение. Но мы должны стремиться к глобальной оптимизации и рассматривать все этапы жизненного цикла информационных моделей (генерацию, хранение, сопровождение, применение) вплоть до их ликвидации или замены. В первую очередь информационные модели геопространства (и топографических поверхностей) создаются для многократного использования. Поэтому задача поставщиков информационных моделей – обеспечить пользователей современным и наиболее эффективным инструментом. Локальная оптимизация процессов создания информационных моделей в рамках топографо-геодезического производства должна выполняться в условиях дополнительных ограничений – требований к содержанию и структуре информационных моделей.

Поэтому имеет смысл перечислить наиболее часто решаемые задачи и рассмотреть возможности использования информационных моделей топографических поверхностей для их решения. Эти задачи следующие:

1) определение «разности» или «суммы» двух поверхностей. Задача возникает при проектировании вертикальной планировки участков местности, когда требуется определить рабочие отметки – разности высот проектируемой и существующей поверхностей. В информационных моделях, предназначенных для использования в навигационных системах, возникает необходимость наложения на топографическую поверхность зданий, сооружений и т. п. Задача вычисления разности двух поверхностей возникает при определении объемов земляных работ. При оценке сложности этих задач нужно иметь в виду, что в общем случае две модельные сетки (регулярные или нерегулярные) не совпадают;

2) отслеживание изолиний и близкая к ней задача определения границ (и площади) затопления при проектировании водохранилищ либо при прогнозировании наводнений;

3) выбор положения проектной линии или поверхности в соответствии с заданной целевой функцией и ограничениями – задача вертикальной планировки территории;

4) построение профилей и определение видимости между точками;

5) определение стока в данной точке. Эта задача, например, возникает при проектировании нефтепроводов. Ее смысл заключается в том, чтобы определить путь, по которому нефть (или другая жидкость) будет растекаться при разрыве трубопровода в данной точке или нескольких таких точках;

6) определения бассейна водосбора в некоторой заданной точке;

7) получение аксонометрической, перспективной или иной проекции поверхности;

8) различного рода картометрические задачи, включая автоматизированное составление легенды на моделируемый участок;

9) проблема генерализации изображения топографической поверхности;

10) объединение или обобщение нескольких цифровых моделей;

11) преобразование системы координат и/или высот;

12) обновление информационных моделей топографических поверхностей с целью поддержания их в актуальном состоянии, то есть, на современном уровне;

13) такие тривиальные и массовые задачи, как вычисление отметок, уклонов и т. п. на множестве определяемых точек.

Почти все перечисленные задачи могут быть сведены либо к вычислению высоты точки по ее плановым координатам, либо к обратной задаче, когда требуется найти точку или все точки по заданной высоте. По затратам процессорного времени регулярные дискретные модели находятся вне всякой конкуренции. Решение тех же задач на нерегулярных дискретных моделях иногда требует времени вычислений на 2–3 порядка больше. Даже для решения элементарной задачи нахождения высоты в точке с заданными координатами в среднем необходимо выполнить проверку принадлежности точки каждому второму треугольнику; в лучшем случае потребуется просмотреть один треугольник, а в худшем – все треугольники. Для ускорения поиска нужного треугольника могут использоваться квадродеревья.

При треугольном разбиении односвязной области моделирования число треугольников N равно

$$N = 2(n-1) - m,$$

где n – число всех исходных точек; m – число исходных точек на границе области триангуляции. Если плотность распределения точек сравнительно равномерна, и область моделирования близка к квадратной, то $m \approx 4\sqrt{n}$. Тогда отношение числа треугольников к числу исходных точек при возрастании числа точек будет стремиться к 2

$$\lim_{n\to\infty}\frac{2(n-1)-m}{n}=2,$$

и при большом числе точек можно считать, что число треугольников примерно равно удвоенному числу точек. Потребность в памяти для представления триангуляции тогда может быть выражена как

$$M=(3K_1+6K_2)n,$$

где K_1 – число единиц памяти, резервируемых под одно значение координаты или высоты; K_2 – число единиц памяти для хранения указателя на вершину треугольника. Для электронных вычислительных машин с байтовой организацией памяти вполне достаточно, если $K_1=4$ и $K_2=2$ байтам.

За единицу измерения числа точек возьмем 1 тысячу точек. Если на каждом треугольном элементе поверхность представляется плоскостью, то для представления треугольников на множестве из 1 000 точек потребуется менее 24 Кбайт, что для современных ЭВМ является вполне приемлемой величиной. Для аппроксимации элемента поверхности уравнением второй степени потребуется уже около 48 Кбайт и т. д. В итоге потребность в оперативной памяти приближается к критической точке. Но поскольку получение аппроксиманта для отдельного элемента не представляет больших трудностей, то целесообразно отказаться от хранения коэффициентов уравнений, описывающих каждый элемент кусочно-непрерывной поверхности, и использовать дискретные модели.

Нерегулярные дискретные информационные модели топографических поверхностей представляются разумным компромиссом между частично противоречивыми требованиями к их точности, используемому объему оперативной памяти и времени (стоимости) их получения. Такие характеристики моделей разного типа, как объем необходимой памяти, есть смысл сравнивать только в том случае, если они обеспечивают одинаковую точность представления поверхности. В одной из работ сообщалось о практических экспериментах по сравнению точности квадратной сетки с сеткой произвольных треугольников. Результаты экспериментов показали, что точность является одинаковой, если число узлов сетки квадратов превышает число узлов триангуляции в 8–10 раз.

Далее будем считать, что регулярная и нерегулярная модели имеют одинаковую точность, если число узлов в первой в 10 раз больше числа узлов во второй. Сравним теперь объемы памяти, необходимые для реализации регулярной и нерегулярной моделей. Из приведенных выше расчетов следует, что в нерегулярных дискретных моделях одна точка порождает потребность в 24 байтах. В моделях на сетке квадратов с увеличением ее размеров отношение числа параметров к числу узлов стремится к 1. Для каждого узла достаточно хранить один параметр – значение высоты. Если для его представления использовать 4 байта, то при фиксированном объеме оперативной памяти число узлов квадратной сетки будет лишь в 6 раз больше числа узлов

нерегулярной модели. Следовательно, точность нерегулярной модели достигнута не будет, и результат не в пользу регулярной модели.

Соотношение коренным образом меняется при переходе к двухбайтовому представлению. Накладными расходами по хранению девяти параметров, характеризующих регулярную дискретную модель

$$H = \{X_0, Y_0, Z_0, \alpha, S_x, S_y, S_z, L_x, L_y, Z\},$$

можно пренебречь. Тогда на той же памяти может быть размещено число узлов квадратной сетки, в 12 раз превышающее число узлов триангуляции, и можно ожидать, что точность регулярной модели будет не хуже.

При реализации любого метода моделирования требуется одновременно размещение в оперативной памяти исходной и конечной (или ее части) моделей. Для сокращения общей потребности в объеме оперативной памяти координаты и высоты исходных точек также могут представляться как целые числа длиной два байта. Если моделирование выполняется отдельно для каждого листа карты или плана, то минимальный шаг дискретизации координат может быть выбран таким, что его величина на чертеже составит менее 0,01 мм. Но это даже излишняя точность. Вполне удовлетворительные результаты будут получены с шагом 0,05 мм или больше, что позволит обрабатывать в одной системе координат не менее 25 соседних трапеций или планшетов.

Такая модель дает возможность представления высот в диапазоне от $H_{\min} = 1$ до $H_{\max} = 65535$ с ошибкой не более 0,5 sh. Теперь нужно оценить эти дополнительные ограничения, для чего уместно задать два вопроса: «какова полоса пропускания модели при измерении высот с максимальной практической точностью», и «каким будет шаг квантования высот (и величина максимальной ошибки), если требуется представление любой точки земной суши или дна океана, т. е. когда полоса пропускания должна быть максимальной?».

При топографических съемках максимальная точность измерения высот составляет 0,01 м, и с помощью такой модели может быть представлен любой участок местности с разностью высот около 1 300 м. Предположение о том, что при решении практических задач на участке топографической поверхности с разностью высот более 1 км может потребоваться точность ее представления порядка 0,01 м, выглядит маловероятным. Следовательно, с разумной точностью может быть представлен любой участок местности ограниченных размеров.

Что касается возможности представления всей земной поверхности в одной модели, то максимальная разность высот и глубин в пределах всего земного шара составляет около 20 км (8 848 м – высота Джомолунгмы и 11 022 м – глубина Марианского желоба в Тихом океане). Следовательно, использование двухбайтового представления требует дискретизации значений высот с шагом 19 870/65 535 = 0,30 м, что позволяет фиксировать высоту любой ее точки с максимальной ошибкой около 0,15 м. Так как существование точек с разностью высот в несколько километров возможно лишь в пределах трапеции мелкого или, самое большее, среднего масштаба, то такая точность более чем

достаточна. Таким образом, двухбайтовое представление высот в дискретных моделях фактически не налагает никаких ограничений на отображаемую топографическую поверхность.

С рассмотренных позиций регулярные дискретные модели обладают преимуществами перед нерегулярными, но несколько проигрывают им в представлении областей со сложными границами. В принципе, возможны три решения этой задачи: топологическое преобразование области моделирования, хранение границ связных участков топографической поверхности и использование значений высот для описания границ.

В последнем случае дополнительная память для хранения границ, как совокупности логических данных, не требуется. Но налагается дополнительное ограничение на дискретизацию высот. Можно принять соглашение, в соответствии с которым высоты в узлах сетки могут быть только положительными. Тогда нулевые значения высот на совокупности смежных узлов являются признаком того, что поверхность в данной области не существует либо при составлении карт не изображается горизонталями. Нулевые значения высот могут служить признаком неопределенности и свидетельствовать о том, что в данной подобласти, определяемой совокупностью узлов с нулевыми высотами, съемка поверхности не выполнялась. Наконец, некоторое значение или значения могут резервироваться для обозначения участков, на которых топографическая поверхность «не существует» (здания и т. п.). Ограничения на квантование высот проявляются в том, что требуется либо уменьшить чувствительность (увеличить шаг квантования по уровню), сохранив прежнюю полосу пропускания, либо поступить наоборот, либо принять некоторый промежуточный вариант, изменив оба параметра.

Но тут просматриваются контуры новой проблемы: «Какой должна быть разрешающая способность регулярной дискретной модели, чтобы обеспечить необходимую точность отображения границ различных участков топографической поверхности». Уменьшая шаг квадратной сетки, можно достичь любой заранее заданной точности. Но при этом возрастает число узлов и возникает вопрос о потребности в оперативной памяти. Основные положения [24] требуют, чтобы ошибки положения четких контуров относительно ближайших точек съемочного обоснования не превышали 0,5 мм или 0,7 мм в горной местности. Погрешности в положении четких контуров относительно друг друга не должны превышать 0,4 мм.

Чисто формально перечисленные требования могут быть удовлетворены присваиванием интервалу между узлами сетки значения около 0,6 мм на чертеже. Побочным положительным эффектом будет повышение точности отображения поверхности. Однако такое решение сопряжено с негативными последствиями и не является удовлетворительным. Если средние размеры трапеции или планшета принять равными 50 × 50 см2 (плюс ширина полосы перекрытия), то необходимо около 106 параметров или 2 Мбайта оперативной памяти. Искажения криволинейных контуров в силу особенностей человеческого восприятия будут практически не заметны, но прямолинейные

контуры с произвольной ориентацией относительно координатных осей (здания и т. д.) будут сильно искажены вследствие проявления «лестничного эффекта».

Следовательно, третий вариант вполне пригоден для решения инженерных задач даже на более крупной сетке, но его реализация для целей картографирования связана с риском: чрезмерные требования к объему оперативной памяти катастрофически сузят круг потенциальных пользователей. Пока что наиболее реален второй способ – хранение границ в виде специфического элемента цифровой модели топографической поверхности.

Можно указать еще один критерий, который обычно игнорируется, – это устойчивость к ошибкам. Информационные модели предназначаются для долговременного хранения, но информация в процессе хранения может искажаться. Изменение только одного коэффициента в уравнении непрерывной поверхности может означать ее полную утрату. В кусочно-непрерывных и нерегулярных дискретных моделях подобные происшествия носят локальный характер и сопровождаются меньшими потерями. В регулярных моделях значительные искажения высот (грубые ошибки) в некоторых узлах могут обнаруживаться и корректироваться программным путем, ошибки восстановления высот и их влияние будут минимальными.

Преимущества регулярных дискретных моделей становятся решающим фактором в проблеме выбора структуры типовой модели, если рассмотреть вопрос об использовании информационных моделей топографических поверхностей. В конце концов, информационные модели земной поверхности создаются для решения тех или иных задач потребителями топографо-геодезической продукции. Уже самим фактом своего первоначального появления они обязаны потребностям проектирования. Если для топографо-геодезического производства информационные модели представляют собой конечную продукцию, то с позиций пользователей они являются лишь средством для достижения их собственных (пользователей) целей.

Завершая анализ достоинств и недостатков различных информационных моделей топографических поверхностей, необходимо остановиться на одном внешнем факторе, оказывающем на экономические показатели информационных моделей доминирующее влияние. Информационную модель можно рассматривать как некоторую систему. Известно, что эффективность функционирования любой системы определяется степенью адаптации системы к ее ближайшему окружению, среде. Для информационных моделей таким окружением служит ЭВМ.

Стоимость автоматизированного составления топографических карт и планов продолжает оставаться высокой. Ее снижение является постоянной заботой разработчиков автоматизированных систем геомоделирования. В работе [13] сообщалось об одном эксперименте по оценке стоимости обработки на ЭВМ различных классов. В качестве теста бралась задача моделирования загрязнения бухты. Оказалось, что время обработки на мини-ЭВМ было в 2–9 раз больше, а стоимость – в 80–90 раз ниже, чем аналогичные показатели на больших машинах. В других экспериментах стоимость снижалась примерно в 300 раз (этот показатель зависит от класса решаемых задач). Как видим, на

эффективность решаемых задач существенно влияет выбор технических средств. В настоящее время нет такого разнообразия ЭВМ, как это было перед появлением персональных компьютеров, и выбор типа ЭВМ довольно ограничен. Как правило, обсуждению подлежат такие характеристики компьютера, как частота процессора и общей шины, объем оперативной памяти и тому подобные параметры.

Можно отметить только, что в свое время среди других преимуществ персональных компьютеров указывалась короткая длина слова (2 байта) и низкая стоимость обработки. Но если последняя сохранилась, то технические характеристики ПЭВМ возросли многократно. В своем развитии персональные компьютеры достигли той точки, когда обычными становятся уже 64-разрядные микропроцессоры. Фирмой AMD такой процессор уже создан. Такие успехи в совершенствовании технических характеристик компьютеров стали возможны благодаря процессу миниатюризации и достижениям микроэлектроники.

Повышение эффективности моделирования может осуществляться минимизацией используемых ресурсов ЭВМ, в частности, использованием операций целочисленной арифметики и укороченной длины слова, то есть двух байт. Время выполнения программ, интенсивно использующих операции над числами с плавающей запятой, больше времени выполнения программ, реализующих те же алгоритмы на целочисленной арифметике. Причина заключается в том, что за единицу времени целочисленных команд выполняется больше, чем команд с вещественной арифметикой. Кроме того, операция деления (умножения) целого числа на число, равное 2^n, может быть заменена операцией арифметического сдвига делимого (первого сомножителя) вправо (влево) на n разрядов, при этом не требуется загрузка второго операнда. Поэтому реализация целочисленных регулярных дискретных моделей может существенно повысить эффективность информационного моделирования топографических поверхностей.

8.10. Методы моделирования топографических поверхностей

Под методом математического моделирования топографической поверхности будем понимать тройку

$$M = (H_1, H_2, T),$$

где H_1 – первичная математическая модель поверхности; H_2 – ее вторичная математическая модель; T – отображение $T : H_1 \to H_2$ первичной модели во вторичную. Иными словами, метод математического моделирования представляет собой совокупность входной и выходной математических моделей и некоторое преобразование (оператор), ставящее в соответствие элементам базового множества модели H_1 элементы из базового множества модели H_2, и отношениям из H_1 – отношения в H_2.

Под первичной математической моделью здесь понимаются исходные данные, обязательным компонентом которых является множество дискретных

точек с известными координатами и высотами. Иногда исходные данные включают в себя сведения о структурных линиях на топографической поверхности. Наличие структурных линий весьма желательно, но не все существующие программные комплексы способны их обрабатывать.

Основной причиной необходимости преобразования первичных моделей является их плохая структурированность или полное отсутствие таковой. Часто исходные данные представляют собой просто набор точек с определенным положением и высотой. Решение любых задач на таких данных неэффективно либо вообще невозможно.

Методом информационного моделирования топографических поверхностей будем называть отображение метода математического моделирования на структуру памяти и структуру управления ЭВМ $M_i : M \to C$.

Для оценки методов моделирования топографических поверхностей их классификацию произведем, прежде всего, по типам множеств, участвующих в отображении. Тогда все методы моделирования можно разбить на четыре класса:

отображение дискретного множества на непрерывное;
отображение дискретного множества на дискретное;
отображение непрерывного множества на непрерывное;
отображение непрерывного множества на дискретное.

Вид преобразований предопределен типом участвующих в отображении множеств и является, следовательно, объективно необходимым. Поэтому нет смысла противопоставлять один класс преобразований другому. Сравнение достоинств и недостатков того или иного вида преобразований целесообразно производить исключительно в рамках одного класса. Композиции преобразований, например, из дискретного множества в непрерывное, а затем из непрерывного в дискретное, рассматриваться не будут в силу их очевидности.

8.11. Методы отображения дискретного множества на непрерывное

При моделировании топографических поверхностей или кривых на них возникает задача восстановления по конечному числу исходных точек функции одной или двух переменных. Первая задача является более простой, и ее решение требуется при построении профилей или горизонталей по заданному набору точек. Восстановление функций двух переменных по заданным точкам – это основная задача при моделировании топографической поверхности.

Методы отображения дискретного множества на непрерывное отличаются наибольшим многообразием по сравнению с другими методами. Главными факторами, оказывающими влияние на выбор конкретного способа и на его свойства, служат:

- целевая функция;
- схема выборки исходных точек;
- способ конструирования поверхности;

– свойства используемых функций моделирования.

Три фактора из четырех перечисленных представлены на рис. 8.43. По типу целевой функции методы отображения дискретного множества на непрерывное подразделяются на *методы интерполирования* и *методы сглаживания,* или *аппроксимации*.

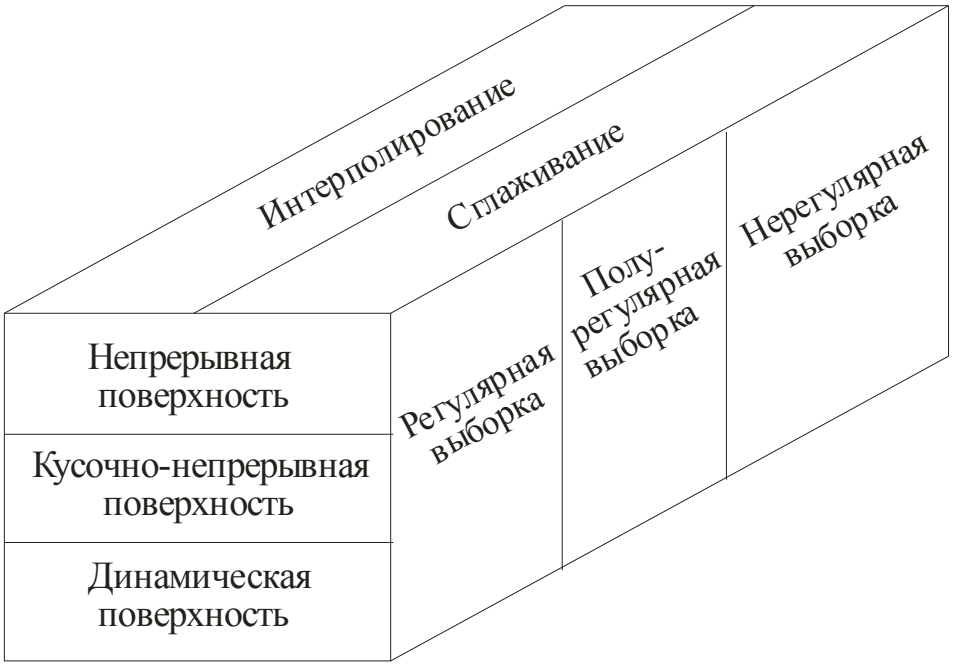


Рис. 8.43. Методы отображения дискретного множества на непрерывное

В зависимости от схемы выборки исходных данных различают методы построения модели по регулярным, полурегулярным и нерегулярным выборкам. Наиболее благоприятными свойствами обладают регулярные выборки. Наибольшие математические трудности возникают при нерегулярных схемах выборки. Но методы, работающие на нерегулярных схемах выборки, применимы также для полурегулярных и регулярных схем. И некоторые из них при этом могут существенно упрощаться.

По способу конструирования поверхности соответствующие им модели разделяют на непрерывные, кусочно-непрерывные и динамические. Классификация методов по типу используемых при моделировании функций не производится, но наибольшей популярностью и наибольшей эффективностью характеризуются методы, в которых в качестве математического аппарата используются алгебраические полиномы.

Задача отображения дискретного множества на непрерывное формулируется следующим образом. Пусть дано конечное множество точек $\{P_i\}_i^n$ с известными значениями высот Z_i. Задача заключается в отыскании функции одной или двух переменных $H(P)$ в соответствии с тем или иным условием или целевой функцией, определенным образом связывающей отыскиваемую функцию с исходными точками. В качестве целевой функции выбирается либо условие точного прохождения поверхности через заданные точки, и тогда возникает задача интерполяции (рис. 8.44), либо условие,

допускающее, но ограничивающее отклонения поверхности от исходных точек, – так называемая задача сглаживания (рис. 8.45). В такой постановке задачи решение неоднозначно, поскольку в качестве моделирующих могут быть выбраны функции различного вида. Поэтому налагаются дополнительные условия или ограничения: указывается тот или иной класс функций, используемых для моделирования, порядок разложения и т. п. На решение задачи оказывает влияние, часто значительное, схема выборки исходных точек.

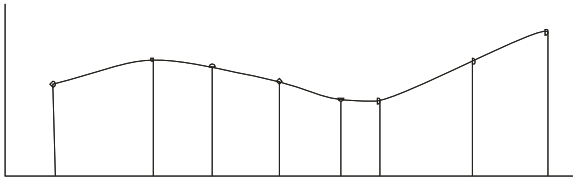
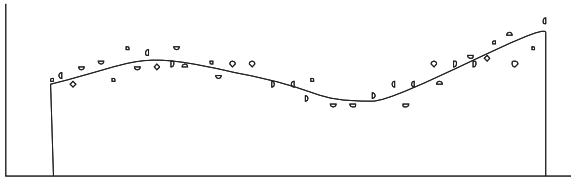
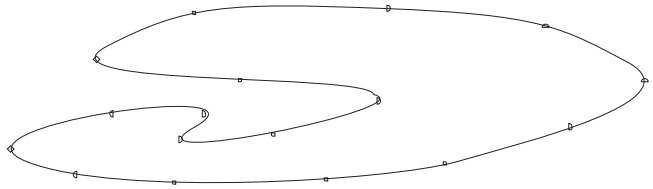


Рис. 8.44. Интерполирование функции f(x)



8.45. Сглаживание функции f(x)

Следует различать задачу восполнения функций одной переменной от задачи моделирования кривых, заданных своими координатами на той или иной поверхности и не являющихся однозначными функциями одной переменной (рис. 8.46). Такими кривыми являются замкнутые и часто даже разомкнутые горизонтали. Кроме того, построение гладких кривых неизбежно возникает не только при вычерчивании горизонталей, но и при автоматическом создании картографических изображений элементов ситуации: гидрографии, растительности и т. п.



8.46. Интерполирование кривой

8.12. Методы интерполирования

Если $H(P)$ – уравнение функции одной или двух переменных, то условие интерполирования означает, что для всех исходных n точек должно выполняться равенство

$$H(P_i) = z_i, \qquad (8.23)$$

где z_i – значение высоты в точке P_i. Условие интерполяции можно записать также в виде соотношения

$$\sum_{i=1}^{n} | h(P_i) - z_i | = 0 \quad (i = 1, ..., n), \qquad (8.24)$$

которое выполнимо, если выполняется предыдущее равенство.

Наибольшее распространение получил такой способ построения интерполяционной функции, называемый линейным интерполированием, когда

функция $H(P)$ рассматривается как линейная комбинация фиксированных функций $h(P)$:

$$H(P) = \sum_{i=1}^{n} c_i h_i(P_j) \quad (j = 1, ..., n), \qquad (8.25)$$

где c_i – некоторые коэффициенты. Так как правая часть представляет собой сумму фиксированных функций, которые при построении функций двух переменных могут рассматриваться как поверхности, то иногда этот метод называют также суммированием поверхностей.

Неизвестные c_i отыскиваются по способу неизвестных коэффициентов, для чего для каждой исходной точки составляется уравнение вида

$$\sum_{i=1}^{n} c_i h_i(P_j) = z_j$$

или

$$\sum_{i=1}^{n} c_i h_i(x_j, y_j) = z_j. \qquad (8.26)$$

В результате получают систему n линейных уравнений с n неизвестными, из решения которой находят значения c_i. Конкретные способы интерполяции отличаются друг от друга выбором фиксированных функций $h(P)$. В математике наиболее изучен случай, когда фиксированные функции являются полиномами. Поэтому первые попытки создания информационных моделей топографических поверхностей были связаны с применением полиномов. Однако практика показала, что такой подход не позволяет добиться удовлетворительных результатов.

Известно, что чем лучше дифференциальные свойства восстанавливаемой функции, тем точнее осуществляется ее аналитическое представление. Топографические поверхности не обладают хорошими дифференциальными свойствами, довольно часто они не являются даже гладкими. Причина подобных неудач (применения полиномов) была объяснена в монографии [29, с. 5]: «Многочлены и рациональные дроби обладают рядом недостатков как аппарат приближения для функций с особенностями и для функций с не слишком большой гладкостью. Основной недостаток состоит в том, что их поведение в окрестности какой-либо точки определяет их поведение в целом». По существу, это приговор многочленам как аппарату представления топографических поверхностей.

Теорема Вейерштрасса утверждает, что для непрерывной на отрезке $[a,b]$ функции $f(x)$ существует такой многочлен $P(x)$, для которого при любом сколь угодно малом $\varepsilon > 0$ во всех точках отрезка выполняется соотношение $|f(x) - P(x)| < \varepsilon$. Поэтому стандартный прием повышения точности интерполирования состоит в увеличении числа исходных точек и повышении

степени интерполяционного полинома. Но, как сообщается в [15], еще в 1901 г. было обнаружено, что, вопреки теореме Вейерштрасса, интерполяционный процесс может расходиться даже для гладких и сколь угодно раз дифференцируемых функций:

$$\lim_{n\to\infty} \max_{a\le x\le b} \left| f(x) - P(x) \right| = \infty.$$

В частности, такой функцией является

$$f(x) = \frac{1}{1+x^2}.$$

При топографических съемках, как правило, нет возможности подбирать положение узлов, поскольку топограф при выполнении работ не может иметь представления о том, каким методом в дальнейшем будет создаваться модель поверхности. Его задача – квалифицированно выполнить съемку земной поверхности, а не пытаться подгонять метод съемки под метод моделирования, что совершенно невозможно. Поэтому при моделировании топографических поверхностей возрастание ошибки представления поверхности может быть устранено не путем подбора узлов интерполяции, а использованием обобщенных многочленов.

Для функции одной переменной обобщенные полиномы имеют вид [15]:

$$H_n(x) = c_0 f_0(x) + c_1 f_1(x) + \ldots + c_n f_n(x). \qquad (8.27)$$

Если в обобщенном полиноме $f_0(x) = 1$, $f_1(x) = \cos x$, $f_2(x) = \sin x$ и т. д., то его называют *тригонометрическим многочленом n-го порядка* и записывают в виде

$$H_n(x) = \frac{a_0}{2} + \sum_{i=1}^{n} (a_i \cos x + b_i \sin x).$$

Если коэффициенты тригонометрического многочлена вычисляются по формулам:

$$a_0 = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x)dx;$$

$$a_i = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x)\cos ixdx;$$

$$b_i = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x)\sin ixdx,$$

то их называют *коэффициентами Фурье*, а многочлен – *рядом Фурье*. О рядах Фурье известно, что они обладают замечательным свойством: среднее квадратическое отклонение

$$\frac{1}{\pi} \int_{-\pi}^{\pi} (f(x) - H_n(x))^2 dx$$

принимает наименьшее по сравнению с другими тригонометрическими многочленами того же порядка значение при любом n. Ряды Фурье при

геомоделировании использовались для построения гладких кривых (профилей или горизонталей) по заданным точкам. В настоящее время они не используются в связи с разработкой более эффективных методов. При создании информационных моделей топографических поверхностей использовались двойные ряды Фурье [10]. Но их применение в указанных целях не дало удовлетворительных результатов.

Американский геодезист Р. Харди предложил [36] для моделирования гладких поверхностей в качестве фиксированных функций использовать *квадратичные функции*, или *квадрики*:

$$h(P_i) = \sqrt{(x - x_i)^2 + (y - y_i)^2 + B}. \qquad (8.28)$$

При $B = 0$ данное уравнение описывает конус, а при $B \neq 0$ – гиперболоид с вертикальной осью. Общее уравнение поверхности представляет собой сумму частных квадрик:

$$H(x, y) = \sum_{i=1}^{n} c_i \sqrt{(x - x_i)^2 + (y - y_i)^2 + B}. \qquad (8.29)$$

В связи с этим данный математический метод представления топографических поверхностей называют *мультиквадриками* и *поликвадриками*. Модель топографической поверхности, полученная с применением мультиквадрик, производит впечатление значительно сглаженной. Приемлемые результаты с использованием мультиквадрик были получены при моделировании спокойного рельефа со сравнительно равномерным расположением исходных точек [17]. Метод мультиквадрик мало пригоден для моделирования топографических поверхностей с резкими формами.

При оценке мультиквадрик как аппарата представления топографических поверхностей приходится учитывать чисто вычислительные аспекты. При вычислении коэффициентов c_i возникает плотно заполненная матрица линейных уравнений. Пусть число исходных точек составляет 10 000. Если коэффициенты уравнения представлять как вещественные числа с удвоенной точностью (8 байт), то для хранения матрицы коэффициентов потребуется 800 Мбайт. На одном номенклатурном листе топографической карты число точек иногда может быть еще больше. Поэтому при достаточно большом числе исходных точек, когда их тысячи или десятки тысяч, в проблему может превратиться как само хранение матрицы (требуется оперативная память для $O(n^2)$ чисел), так и время вычислений, пропорциональное кубу от числа исходных точек, – $O(n^3)$.

В статье [20] отмечалось, что результаты интерполяции методом мультиквадрик зависят от выбора поверхности тренда. Кроме того, в других работах были даны взаимно противоречивые рекомендации по выбору параметра *B*. С формальных позиций достаточно очевидно, что его увеличение ухудшает обусловленность системы линейных уравнений. Если коэффициент *B* достаточно большой по сравнению с расстояниями между точками, то коэффициенты системы линейных уравнений при этом становятся все более

«похожими», происходит их выравнивание, и определитель системы стремится к нулю. Из-за ограниченного числа разрядов для представления чисел в ЭВМ и ошибок округления результаты вычисления коэффициентов c_i при большом числе уравнений получаются недостоверными. Было установлено, что при неравномерном распределении исходных точек может происходить потеря 4–5 знаков при решении системы уравнений.

Численные эксперименты, выполнявшиеся в свое время на ВЦ СО АН СССР, показали, что при использовании арифметики с обычной точностью (4 байта) относительно надежно значения c_i определяются, если число уравнений не превышает 50–70. Применение арифметики с двойной точностью приводит к увеличению в два раза объема оперативной памяти и к возрастанию времени вычислений. Поэтому применение метода мультиквадрик для сложных топографических поверхностей, представленных большим числом исходных точек, не реально.

В качестве «национальной особенности» мультиквадрик можно отметить тот факт, что в Советском Союзе было защищено 3 или даже 4 кандидатских диссертации по их применению для моделирования топографических поверхностей.

Еще одним способом восстановления непрерывной поверхности является ее представление в виде среднего весового:

$$H(x,y)=\frac{\sum_{i=1}^{n} p_i z_i}{\sum_{i=1}^{n} p_i}, \qquad (8.30)$$

где вес p – некоторая положительная убывающая функция от расстояния [9]. Хотя данный способ дает гладкую поверхность, целесообразность его применения сомнительна. Тем не менее, попытки использования способа среднего взвешенного не столь малочисленны, как этого можно было бы ожидать. Конкретные реализации данного метода отличаются, в основном, способом определения весовых функций. Привлекательная сторона метода среднего весового состоит в том, что при его использовании не требуется решать системы уравнений, и значение высоты поверхности может быть получено в любой ее точке непосредственным применением формулы (8.30). Особенность данного метода состоит также в том, что в зависимости от определения весовой функции он является либо интерполяцией, либо аппроксимацией.

Наиболее интересным и теоретически обоснованным в данной группе представляется метод, являющийся аналитическим решением задачи *сплайн-интерполяции* в областях с произвольно расположенными исходными точками [7]. Он предложен сравнительно недавно, уже после разработки сплайнов на подпространствах; по некоторым источникам, этот метод использовался в гидрографических системах.

Аналитическая сплайн-интерполяция отличается от других методов данного класса уже постановкой задачи. В рассматривавшихся выше методах в качестве дополнительных ограничений указывается конкретный вид аналитического выражения. Одни разработчики выбирают алгебраические полиномы, другие – тригонометрические многочлены, например, двойные ряды Фурье, третьи – мультиквадрики и т. д. В сущности, каждый такой выбор является актом волевого решения.

В методе сплайн-интерполяции делаются разумные предположения о дифференциальных свойствах восстанавливаемой функции, и указывается класс функций, на котором отыскивается решение – вид аналитического выражения. В частности, при решении задачи интерполяции функции двух переменных предполагается, что моделируемая поверхность обладает свойством минимальной кривизны:

$$\Phi(u) = \iint_D \left[\left(\frac{\partial^2 u}{\partial x^2} \right)^2 + 2 \left(\frac{\partial^2 u}{\partial x \partial y} \right)^2 + \left(\frac{\partial^2 u}{\partial y^2} \right)^2 \right] dxdy = \min, \qquad (8.31)$$

и решение отыскивается среди всех функций класса ω_2^2, то есть среди всех функций от двух переменных, имеющих непрерывную вторую производную. Решением сформулированной задачи является *сплайн-функция* [7]:

$$\sigma(x, y) = \frac{1}{2} \sum_{i=1}^{n} \lambda_i a_i + \nu_0 + \nu_1 x + \nu_2 y, \qquad (8.32)$$

где n – число исходных точек; $d_i = (x - x_i)^2 + (y - y_i)^2$; $a_i = d_i \ln d_i$.

Коэффициенты λ_i, ν_0, ν_1 и ν_2 находят из решения системы

$$\begin{pmatrix} 0 & a_{12} & \dots & a_{1n} & 1 & x_1 & y_1 \\ a_{21} & 0 & \dots & a_{2n} & 1 & x_2 & y_2 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & 0 & 1 & x_n & y_n \\ 1 & 1 & \dots & 1 & 0 & 0 & 0 \\ x_1 & x_2 & \dots & x_n & 0 & 0 & 0 \\ y_1 & y_2 & \dots & y_n & 0 & 0 & 0 \end{pmatrix} \times \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \dots \\ \lambda_n \\ \nu_0 \\ \nu_1 \\ \nu_2 \end{pmatrix} = \begin{pmatrix} \mathrm{Z}_1 \\ \mathrm{Z}_2 \\ \dots \\ \mathrm{Z}_n \\ 0 \\ 0 \\ 0 \end{pmatrix}. \qquad (8.33)$$

Данная система не вырождена, если среди исходных точек хотя бы три точки не лежат на одной прямой. Некоторые особенности (они обсуждались в [7]) системы уравнений и уравнения, описывающего поверхность, требуют аккуратной реализации данного метода моделирования. При возрастании порядка системы наблюдается тенденция к ухудшению ее обусловленности. Метод был рекомендован к использованию, когда число исходных точек не превышает 150.

Во всех перечисленных методах данной группы постулируются возможность представления всей поверхности единым уравнением и ее гладкость. Однако, как уже отмечалось, топографическая поверхность далеко не

всегда является гладкой. Кроме того, все попытки представления достаточно больших участков земной поверхности единым уравнением неизбежно сталкиваются с чисто вычислительными проблемами – необходимостью решения систем уравнений высокого и даже очень высокого порядка, когда число неизвестных может превышать 10^4.

Общим преимуществом методов данной группы является независимость от конфигурации области моделирования и от схемы выборки исходных точек. В случае регулярных схем выборки в некоторых методах формулы могут быть упрощены, или даже получены явные выражения для вычисления коэффициентов уравнений (без решения системы линейных уравнений).

Общим недостатком перечисленных методов (за исключением метода среднего весового) при произвольном расположении опорных точек является необходимость решения больших и очень больших систем линейных уравнений с плотно заполненными матрицами и невозможность отображения негладких поверхностей.

Методы построения непрерывных моделей служат основой конструирования кусочно-непрерывных топографических поверхностей.

8.13. Методы конструирования кусочно-непрерывных поверхностей

Использование кусочно-непрерывного способа моделирования топографических поверхностей имеет своей целью избавление от вычислительных проблем и повышение точности моделирования.

Наибольшие удобства для построения гладкой поверхности возникают, если точки заданы в узлах прямоугольной или квадратной сетки [4]. С восстановлением функций, заданных своими значениями в узлах регулярной сетки, связано появление сплайн-функций.

8.14. Сплайн-функции

Но прежде, чем перейти к моделированию поверхностей, рассмотрим кусочно-непрерывное представление функций одной переменной. Такое представление началось с аппарата сплайнов, предложенного в 1946 г. американским математиком Шенбергом. Своим происхождением и названием сплайны обязаны гибким стержням, которые чертежники применяли при вычерчивании гладких кривых. Такой стержень закреплялся в заданных точках, а затем по нему проводилась, как по линейке, гладкая кривая. В сопромате доказано, что если упругий стержень закрепить в двух точках и в этих точках придать ему определенные направления, то при некоторых упрощающих предположениях кривая оси стержня будет описываться полиномом третьей степени.

В общем случае *сплайном* называют кусочно-непрерывную функцию, обладающую непрерывными производными до n-го порядка включительно. Наибольшее распространение получили *полиномиальные сплайны*, названные так по той причине, что в качестве кусочно-непрерывных функций используются алгебраические полиномы. Их формальное определение состоит в следующем [29].

Пусть на отрезке $[a,b]$ задана сетка

$$\Delta: a = x_0 < x_1 < ... < x_n = b, \qquad (8.34)$$

P – множество полиномов степени не выше m ($m > 0$) и $C^{(k)} = C^{(k)}[a,b]$ – множество непрерывных на $[a,b]$ функций, имеющих непрерывную k-ю производную. Тогда *полиномиальным сплайном* степени m дефекта k называют функцию $\sigma_m(x)$, обладающую свойствами:

1) $\sigma_m(x) \in P_m$ при $x \in [x_i, x_{i+1}]$;

2) $\sigma_m(x) \in C^{m-k}[a,b]$.

Говорят, что сплайн-функция $\sigma_m(x)$ имеет *дефект* k, если $(m - k + 1)$-я производная претерпевает разрыв.

Из полиномиальных сплайнов наибольшей популярностью пользуются сплайн-функции нечетных степеней, прежде всего – линейные и кубические сплайны, что объясняется их практичностью и замечательными математическими свойствами. Оказывается, что для кубического сплайна выполняется так называемое *интегральное соотношение*

$$\int_a^b |f''(x)|^2 dx = \int_a^b |\sigma''(x)|^2 dx + \int_a^b |f''(x) - \sigma''(x)|^2 dx. \qquad (8.35)$$

Кроме того, среди всех функций $f(x)$, проходящих через заданные точки и имеющих на $[a,b]$ непрерывную вторую производную, только для кубического сплайна имеет место соотношение

$$\int_a^b |f''(x)|^2 dx = \min. \qquad (8.36)$$

Так как этот интеграл является хорошим приближением интеграла от квадрата кривизны, то данное свойство называют *свойством минимальной кривизны*.

Пусть на вещественном отрезке $[a,b]$ имеется сетка (8.34), в узлах которой задана своими значениями функция $f(x_i)$ $(i = 0,1,...,n)$. *Кубической сплайн-функцией* называют функцию $S_3(x)$, непрерывную на всем отрезке $[a,b]$ вместе со своей первой и второй производными и совпадающую на каждом отрезке $[x_i, x_{i+1}]$ $(i = 0,1,...,n-1)$ с кубическим полиномом

$$P_i(x) = \sum_{k=0}^{3} a_k^{(i)} (x - x_i)^k \qquad (8.37)$$

и удовлетворяющую следующим условиям:

1) в каждом узле сетки x_i выполняется равенство

$$S_3(x_i) = f(x_i); \qquad (8.38)$$

2) на концах отрезка $[a,b]$ определено одно из граничных условий:

а) $S_3'(a) = f'(a)$, $S_3'(b) = f'(b)$; (8.39.1)

б) $S_3''(a) = f''(a),\ S_3''(b) = f''(b).$ (8.39.2)

Были доказаны существование и единственность такой функции, а также то, что она минимизирует функционал

$$\Phi(u) = \int_a^b [u''(x)]^2\, dx, \qquad (8.40)$$

где

$$u \in W_2^2[a, b],\ u(x_i) = f(x_i)\ \ (i = 0,1, ..., n). \qquad (8.41)$$

Для нахождения аналитического представления сплайн-функции можно воспользоваться тем, что ее вторая производная непрерывна и линейна на каждом отрезке $[x_i, x_{i+1}]$ $(i = 0,1, ..., n-1)$ сетки Δ. Следовательно, можно написать

$$S_3''(x) = a_2^{(i-1)} \frac{x_i - x}{d_i} + a_2^{(i)} \frac{x - x_{i-1}}{d_i}, \qquad (8.42)$$

где

$$d_i = x_i - x_{i-1}, \qquad (8.43)$$

$$a_2^{(i)} = S_3''(x_i). \qquad (8.44)$$

Проинтегрировав дважды обе части равенства (8.42), получим

$$S_3(x) = a_2^{(i-1)} \frac{(x_i - x)^3}{6d_i} + a_2^{(i)} \frac{(x - x_{i-1})^3}{6d_i} + B_i \frac{x_i - x}{d_i} + C_i \frac{x - x_{i-1}}{d_i}, \qquad (8.45)$$

где B_i, C_i – константы интегрирования. Для их определения подставим $x = x_{i-1}$ и $x = x_i$ в уравнение (8.45). Тогда с учетом условия (8.38) получим

$$\left.\begin{aligned} a_0^{(i-1)} &= a_2^{(i-1)} \frac{d_i^2}{6} + B_i \\ a_0^{(i)} &= a_2^{(i)} \frac{d_i^2}{6} + C_i \end{aligned}\right\}, \qquad (8.46)$$

где

$$a_0^{(i)} = f(x_i).$$

Определим константы B_i и C_i из уравнений (8.46) и подставим их в (8.45). В результате получим нужное аналитическое представление сплайн-функции через ее вторые производные

$$S_3(x) = a_2^{(i-1)} \frac{(x_i - x)^3}{6d_i} + a_2^{(i)} \frac{(x - x_{i-1})^3}{6d_i} +$$

$$+ (a_0^{(i-1)} - a_2^{(i-1)} \frac{d_i^2}{6}) \frac{x_i - x}{d_i} + (a_0^{(i)} - a_2^{(i)} \frac{d_i^2}{6}) \frac{x - x_{i-1}}{d_i}. \qquad (8.47)$$

Из данного соотношения можно получить выражение для первой производной

$$S_3'(x) = -a_2^{(i-1)}\frac{(x_i - x)^2}{2d_i} + a_2^{(i)}\frac{(x - x_{i-1})^2}{2d_i} +$$

$$+\frac{a_0^{(i)} - a_0^{(i-1)}}{d_i} - a_0^{(i)} - \frac{a_2^{(i)} - a_2^{(i-1)}}{6}d_i. \qquad (8.48)$$

В каждом узле сетки односторонние пределы равны

$$\left.\begin{aligned} S_3'(x-0) &= -a_2^{(i-1)}\frac{d_i}{6} + a_2^{(i)}\frac{d_i}{3} + \frac{a_0^{(i)} - a_0^{(i-1)}}{d_i} \\ S_3'(x+0) &= -a_2^{(i)}\frac{d_i}{3} - a_2^{(i+1)}\frac{d_{i+1}}{6} + \frac{a_0^{(i+1)} - a_0^{(i)}}{d_{i+1}} \end{aligned}\right\}. \qquad (8.49)$$

Так как первая и вторая производные непрерывны, то, приравнивая правые части уравнений (8.49) в каждом узле, получим систему $n - 1$ уравнений для всей сетки

$$\frac{d_i}{6}a_2^{(i-1)} + \frac{d_i + d_{i+1}}{3}a_2^{(i)} + \frac{d_{i+1}}{6}a_2^{(i+1)} = \frac{a_0^{(i+1)} - a_0^{(i)}}{d_{i+1}} - \frac{a_0^{(i)} - a_0^{(i-1)}}{d_i}. \qquad (8.50)$$

Умножим теперь обе части соотношения (8.50) на $\frac{6}{d_i + d_{i+1}}$ и введем обозначения

$$\left.\begin{aligned} \lambda_i &= \frac{d_i}{d_i + d_{i+1}} \\ \mu_i &= \frac{d_{i+1}}{d_i + d_{i+1}} \end{aligned}\right\}. \qquad (8.51)$$

В итоге получим систему из $n - 1$ уравнений, в которой число неизвестных равно $n + 1$

$$\lambda_i a_2^{(i-1)} + 2a_2^{(i)} + \mu_i a_2^{(i+1)} = 6\mu_i\frac{a_0^{(i+1)} - a_0^{(i)}}{d_{i+1}^2} - 6\lambda_i\frac{a_0^{(i)} - a_0^{(i-1)}}{d_i^2}. \qquad (8.52)$$

Для определения двух неизвестных $a_2^{(0)}$ и $a_2^{(n)}$ необходимо задать два дополнительных граничных условия. Наиболее простыми такими условиями являются

$$a_2^{(0)} = 0 \text{ и } a_2^{(n)} = 0, \qquad (8.53)$$

но они, как правило, отличаются от действительных значений. Лучше краевые условия получить из уравнения (8.49). На основании (8.39) можно записать

$$\left.\begin{aligned} f'(a) = a_1^{(0)} &= -a_2^{(0)}\frac{d_1}{3} - a_2^{(1)}\frac{d_1}{6} + \frac{a_0^{(1)} - a_0^{(0)}}{d_1} \\ f'(b) = a_1^{(n)} &= -a_2^{(n-1)}\frac{d_n}{6} + a_2^{(n)}\frac{d_n}{3} + \frac{a_0^{(n)} - a_0^{(n-1)}}{d_n} \end{aligned}\right\}$$

или

$$\left.\begin{aligned} 2a_2^{(0)} + a_2^{(1)} &= \frac{6}{d_1}\left(\frac{a_0^{(1)} - a_0^{(0)}}{d_1} - a_1^{(0)}\right) \\ a_2^{(n-1)} + 2a_2^{(n)} &= \frac{6}{d_n}\left(a_1^{(n)}\frac{a_0^{(n)} - a_0^{(n-1)}}{d_n}\right) \end{aligned}\right\}. \qquad (8.54)$$

Обозначив правые части (8.52) и (8.54) как r_i, получим следующую систему уравнений для определения неизвестных $a_2^{(i)}$

$$\left.\begin{aligned} 2a_2^{(0)} + a_2^{(1)} &= r_0 \\ \lambda_1 a_2^{(0)} + 2a_2^{(1)} + \mu_1 a_2^{(2)} &= r_1 \\ \cdots & \\ \lambda_{n-1} a_2^{(n-2)} + 2a_2^{(n-1)} + \mu_{n-1} a_2^{(n)} &= r_{n-1} \\ a_2^{(n-1)} + 2a_2^{(n)} &= r_n \end{aligned}\right\}. \qquad (8.55)$$

Для определения r_0 и r_n необходимо иметь значения первых производных $a_1^{(0)}$ и $a_1^{(n)}$ на концах отрезка $[a, b]$. Они могут определяться одним из следующих способов:

1) $a_1^{(0)} = 0$, $a_1^{(n)} = 0$;

2) если известны дополнительные точки (i = -1) и (i = n + 1) на концах отрезка, то в виде

$$a_1^{(0)} = \frac{a_0^{(1)} - a_0^{(-1)}}{x_i - x_{-1}};\ a_1^{(n)} = \frac{a_0^{(n+1)} - a_0^{(n-1)}}{x_{n+1} - x_{n-1}}; \qquad (8.56)$$

3) и как

$$a_1^{(0)} = \frac{a_0^{(1)} - a_0^{(0)}}{d_1};\ a_1^{(n)} = \frac{a_0^{(n)} - a_0^{(n-1)}}{d_n} \qquad (8.57)$$

либо их комбинацией.

Если известны дополнительные точки, то общими являются граничные условия

$$\left.\begin{aligned} 2a_2^{(0)} + \mu_0 a_2^{(1)} &= r_0 \\ \lambda_n a_2^{(n-1)} + 2a_2^{(n)} &= r_n \end{aligned}\right\}. \qquad (8.58)$$

Используя данные граничные условия, представим систему уравнений (8.55) в матричной форме

$$\begin{bmatrix} 2 & \mu_0 & 0 & \dots & 0 & 0 & 0 \\ \lambda_1 & 2 & \mu_1 & \dots & 0 & 0 & 0 \\ 0 & \lambda_2 & 2 & \dots & 0 & 0 & 0 \\ & \dots & & \dots & & \dots & \\ 0 & 0 & 0 & \dots & 2 & \mu_{n-2} & 0 \\ 0 & 0 & 0 & \dots & \lambda_{n-1} & 2 & \mu_{n-1} \\ 0 & 0 & 0 & \dots & 0 & \lambda_n & 2 \end{bmatrix} \begin{bmatrix} a_2^{(0)} \\ a_2^{(1)} \\ a_2^{(2)} \\ \dots \\ a_2^{(n-2)} \\ a_2^{(n-1)} \\ a_2^{(n)} \end{bmatrix} = \begin{bmatrix} r_0 \\ r_1 \\ r_2 \\ \dots \\ r_{n-2} \\ r_{n-1} \\ r_n \end{bmatrix}. \qquad (8.59)$$

Полученную систему уравнений вместе с соотношениями (8.47) называют кубической сплайн-функцией II типа. В различных приложениях более удобным оказывается использование сплайн-функций I типа, когда они выражаются через значения первых производных.

На основании определения кубического сплайна для узла x_i можно записать

$$a_0^{(i)} = a_0^{(i-1)} + a_1^{(i-1)} d_i + a_2^{(i-1)} d_i^2 + a_3^{(i-1)} d_i^3. \qquad (8.60)$$

Тогда первая и вторая производные будут иметь вид

$$a_1^{(i)} = a_1^{(i-1)} + 2a_2^{(i-1)} d_i + 3a_3^{(i-1)} d_i^2; \qquad (8.61)$$

$$a_2^{(i)} = 2a_2^{(i-1)} + 6a_3^{(i-1)} d_i. \qquad (8.62)$$

Из соотношений (8.60), (8.61) и (8.62) можно получить выражения для значений коэффициентов $a_2^{(i)}$ и $a_3^{(i)}$. Для этого вначале из (8.60) и (8.61) можно исключить $a_2^{(i-1)}$:

$$a_3^{(i-1)} = \frac{1}{d_i^3}[2(a_0^{(i-1)} - a_0^{(i)}) + d_i(a_1^{(i-1)} + a_1^{(i)})]. \qquad (8.63)$$

Теперь в выражение (8.62) подставим значение $a_2^{(i-1)}$ из (8.61)

$$a_2^{(i)} = \frac{1}{d_i}(a_1^{(i)} - a_1^{(i-1)}) + 3a_3^{(i-1)} d_i. \qquad (8.64)$$

Из уравнения (8.60) следует

$$a_2^{(i-1)} = \frac{1}{d_i^2}(a_0^{(i)} - a_0^{(i-1)}) - \frac{1}{d_i} a_1^{(i-1)} + d_i a_3^{(i-1)}. \qquad (8.65)$$

Аналогичное выражение можно записать для $a_2^{(i)}$:

$$a_2^{(i)} = \frac{1}{d_{i+1}^2}(a_0^{(i+1)} - a_0^{(i)}) - \frac{1}{d_{i+1}} a_1^{(i)} + d_{i+1} a_3^{(i)}. \qquad (8.66)$$

Правые части выражений (8.64) и (8.66) можно приравнять, заменив $a_3^{(i-1)}$ и $a_3^{(i)}$ их значениями из (8.63). Тогда получим соотношение

$$\frac{1}{d_i} a_1^{(i-1)} + 2(\frac{1}{d_i} + \frac{1}{d_{i+1}}) + \frac{1}{d_{i+1}} a_1^{(i+1)} = 3(\frac{a_0^{(i)} - a_0^{(i-1)}}{d_i^2} + \frac{a_0^{(i+1)} - a_0^{(i)}}{d_{i+1}^2})$$

.(8.67)

Используя обозначения (8.51), данное уравнение можно записать как

$$\mu_i a_i^{(i-1)} + 2a_1^{(i)} + \lambda_i a_1^{(i+1)} = 3(\mu_i \frac{a_0^{(i)} - a_0^{(i-1)}}{d_i} + \lambda_i \frac{a_0^{(i+1)} - a_0^{(i)}}{d_{i+1}}). \quad (8.68)$$

Обозначим правую часть данного равенства через c_i и введем граничные условия общего вида

$$\left.\begin{aligned} 2a_1^{(0)} + \lambda_0 a_0^{(1)} &= c_0 \\ \mu_n a_1^{(n-1)} + 2a_1^{(n)} &= c_n \end{aligned}\right\}. \qquad (8.69)$$

Тогда мы приходим к системе линейных уравнений

$$\left.\begin{aligned} 2a_1^{(0)} + \lambda_0 a_1^{(1)} &= c_0 \\ \mu_1 a_1^{(0)} + 2a_1^{(1)} + \lambda_1 a_1^{(2)} &= c_1 \\ \dots \quad\quad\quad\quad\quad &\dots \quad \dots \\ \mu_{n-1} a_1^{(n-2)} + 2a_1^{(n-1)} + \lambda_{n-1} a_1^{(n)} &= c_{n-1} \\ \mu_n a_1^{(n-1)} + 2a_1^{(n)} &= c_n \end{aligned}\right\}, \qquad (8.70)$$

где неизвестными являются значения коэффициентов $a_1^{(i)}$.

Если на концах отрезка существуют дополнительные точки с координатами $(x_{-1}, f(x_{-1}))$ и $(x_{n+1}, f(x_{n+1}))$, то значения всех неизвестных $a_1^{(i)}$ $(i = 0,1,\dots,n)$ могут быть получены непосредственным решением системы уравнений (8.70). В противном случае значения $a_1^{(0)}$ и $a_1^{(n)}$ могут быть получены одним из указанных выше способов, а остальные значения $a_1^{(i)}$ $(i = 1,2,\dots,n-1)$ – из решения системы (8.70).

Чтобы получить представление кубической сплайн-функции через значения коэффициентов $a_1^{(i)}$, заменим в (8.37) коэффициенты $a_2^{(i)}$ и $a_3^{(i)}$ их выражениями для $a_1^{(i)}$ в (8.63) и (8.64). После некоторых преобразований получим представление кубического сплайна на отрезке $[x_{i-1}, x_i]$:

$$S_3(x) = P(x) = a_1^{(i-1)} \frac{(x_i - x)^2 (x - x_{i-1})}{d_i^2} - a_1^{(i)} \frac{(x - x_{i-1})^2 (x_i - x)}{d_i^2} +$$

$$+ a_0^{(i-1)} \frac{(x_i - x)^2 [2(x - x_{i-1}) + d_i]}{d_i^3} + a_0^{(i)} \frac{(x - x_{i-1})^2 [2(x_i - x) + d_i]}{d_i^3}. \quad (8.71)$$

Для решения подобных систем уравнений (8.55) и (8.70) с трехдиагональной матрицей коэффициентов существует эффективный «*метод прогонки*», суть которого заключается в следующем. Пусть имеется система линейных уравнений с трехдиагональной матрицей коэффициентов:

$$\left.\begin{array}{ll} b_1 x_1 + c_1 x_2 & = r_1 \\ a_2 x_1 + b_2 x_2 + c_2 x_3 & = r_2 \\ \ldots & \ldots \ldots \\ a_{n-1} x_{n-2} + b_{n-1} x_{n-1} + c_{n-1} x_n & = r_{n-1} \\ a_n x_{n-1} + b_n x_n & = r_n \end{array}\right\}. \quad (8.72)$$

Эта система уравнений решается за два прохода. При прямом проходе для определения «прогоночных» коэффициентов используются формулы

$$\left.\begin{array}{c} p_k = a_k q_{k-1} + b_k; \\ q_k = -\dfrac{c_k}{p_k}; \\ u_k = \dfrac{r_k - a_k u_{k-1}}{p_k} \ (\mathrm{k} = 1, 2, \ldots, \mathrm{n}); \\ u_0 = 0; \ q_0 = 0. \end{array}\right\} \quad (8.73)$$

Последний коэффициент (при $k = n$) u_k есть неизвестное x_n, то есть $x_n = u_n$. При обратном проходе значения неизвестных вычисляются по формулам

$$x_k = q_k x_{k+1} + u_k \ \ (k = n-1, n-2, \ldots, 1). \quad (8.74)$$

Трехдиагональные матрицы, подобные (8.59), характеризуются строгим диагональным преобладанием, поскольку $\lambda_i + \mu_i = 1$. Поэтому получаемое решение отличается большой устойчивостью и значения неизвестных вычисляются однозначно. Для равномерных сеток уравнения (8.52) и (8.70) сильно упрощаются, что позволяет получить для них явные формулы для вычисления $a_1^{(i)}$ и $a_2^{(i)}$, то есть без решения системы уравнений.

Изложенный вариант кубических сплайн-функций можно считать классическим [1]. Интерес к ним резко возрос после 1964 г. За время существования теории сплайн-функций число публикаций в мире по этой проблеме, вероятно, составляет несколько тысяч.

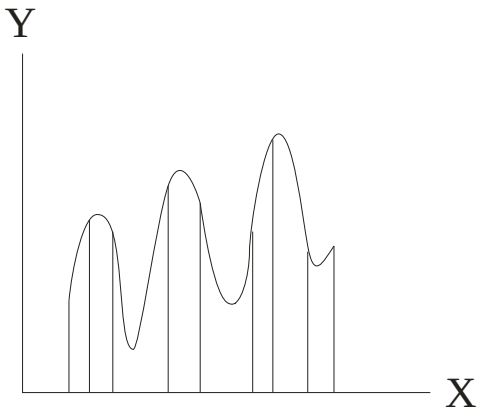


Рис. 8.47. Осциллирующая функция

Первоначально сплайны были разработаны как средство борьбы с двумя проблемами, возникающими при использовании в качестве интерполянтов полиномов высокого порядка. Они были предложены, во-первых, чтобы избавиться от необходимости решать большие системы линейных уравнений с плотно заполненными матрицами коэффициентов; во-вторых, чтобы избежать нежелательных осцилляций, часто возникающих при применении полиномов (рис. 8.47), когда интерполирующая функция сильно колеблется, хотя и проходит через заданные точки.

Дальнейшее расширение теории сплайнов было связано с их обобщением на случай функций многих переменных (прежде всего – билинейные и бикубические сплайны), и сегодня сплайны применяются в самых разнообразных приложениях. По этому поводу отмечалось: «Способ приближения сплайнами интересен прежде всего отношением к нему. Одни считают его универсальным методом решения проблем, стоящих перед численным анализом, и ищут применения ему в самых различных направлениях, другие рассматривают его как очередную дань переменчивой математической моде. По-видимому, истина проходит где-то посередине; в настоящее время область применения этого метода непрерывно растет» [3, с. 256].

Развитие теории сплайн-функций шло по следующим главным направлениям:

– обобщению аппарата сплайнов на функции многих переменных;

– разработке теории локальных сплайнов;

– применению в качестве интерполирующих функций рациональных дробей, тригонометрических полиномов и т. п.;

– использованию сплайнов различной степени и различной степени гладкости.

По сравнению со всеми другими способами бикубические сплайны обеспечивают наибольшую точность представления, тем не менее, как аппарат моделирования топографических поверхностей, они обладают рядом недостатков. Прежде всего, следует признать, что между тонким и упругим стержнем (кубическим сплайном) и топографической поверхностью мало общего. Стержень обладает более удобными свойствами: более гладок, его поведение более предсказуемо. Топографические поверхности такими свойствами обладают далеко не всегда. Кроме того, при использовании сплайнов, как и в других случаях использования полиномов, форма кривой также зависит от выбора системы координат.

В некоторых случаях между исходными точками появляются излишние осцилляции или точки нулевой кривизны. Для борьбы с подобными эффектами Д. Швайкертом были предложены *сплайны в напряженном состоянии* [38]. Если пользоваться механической аналогией, то сплайн в напряженном состоянии – это стержень, к концам которого приложены растягивающие усилия (рис. 8.48). Очевидно, что при изменении растягивающего усилия форма кривой будет изменяться. Однако это предложение для случая хаотично расположенных точек, вероятно, не было реализовано, скорее всего, из-за трудностей вычислительного характера.

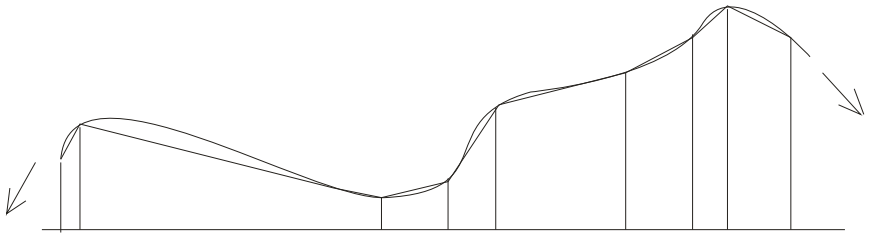


Рис. 8.48. Сплайн в напряженном состоянии

Кроме того, сплайны в напряженном состоянии вызывают определенные трудности и на содержательном уровне. Возможно, что в других приложениях из физических соображений можно установить величину растягивающего усилия. Но при моделировании топографической поверхности, например, при построении профиля по заданным точкам, совершенно не ясно, какое растягивающее усилие нужно прилагать. Может также оказаться, что на одном участке кривой (профиля или горизонтали) требуется одно усилие, а на другом – другое.

Для нас наибольший интерес представляют обобщения сплайнов на случай функций двух переменных и локальные сплайны.

8.15. Локальные сплайн-функции одной переменной

Как отмечалось выше, одной из основных причин разработки методов кусочно-непрерывного представления функций было стремление избежать нежелательных осцилляций интерполирующей функции. Практика применения кубических сплайн-функций со временем показала, что решить эту проблему удается не всегда. Вблизи «плохих» узлов возможно появление излишних осцилляций функции, а также точек нулевой кривизны.

Подобные осцилляции возникают в случаях, когда в одном из соседних узлов первая производная имеет большое значение, а в другом – существенно меньшее (рис. 8.49). Обычно такой эффект наблюдается, если длины двух соседних отрезков сильно отличаются, и он тем больше, чем больше эта разница. Так, на рис. 8.49 длина отрезка $[x_1, x_2]$ намного меньше длины отрезка $[x_2, x_3]$. Именно

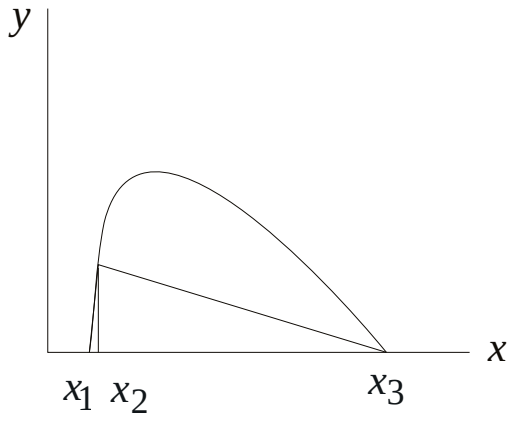


Рис. 8.49. Большое значение f'

вследствие этой разницы в длинах мы наблюдаем всплеск кривой на отрезке $[x_2, x_3]$. Если бы данная кривая была участком профиля земной поверхности, то при «ручной» интерполяции кривую провели бы так, что узел x_2 являлся бы точкой локального экстремума.

Из приведенного примера следует, что весьма желательным является равномерное распределение узлов. Однако при съемках топографической поверхности или профиля отбору подлежат в первую очередь характерные точки. Поэтому следует изменять не методы съемки, а совершенствовать методы моделирования.

Тот же рис. 8.49 подсказывает способ решения задачи – использование значений первой производной в узлах. Изменяя значение первой производной в узле x_2, например, приняв $f'(x_2) = 0$, можно добиться нужного нам эффекта.

В результате подобных рассуждений мы приходим к понятию локального сплайна. Значения производных в опорных точках могут быть заданы в качестве исходных данных или вычислены каким-либо произвольным образом. Тогда для определения интерполирующей функции не требуется решать большие системы линейных уравнений. Сплайн, строящийся по значениям функции и ее производной, называют локальным кубическим сплайном в отличие от определенного выше, который иногда называют полным кубическим сплайном, чтобы подчеркнуть его отличие от локального.

В своем большинстве локальные сплайны основаны на так называемой интерполяции Эрмита, пример которой представлен на рис. 8.50. Пусть исходными данными для интерполяции служат значения функции и ее первой производной в каждом узле интерполяции и на каждом отрезке интерполируемая функция представляется полиномом третьей степени.

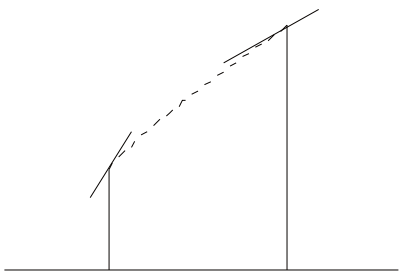


Рис. 8.50. Интерполяция Эрмита

Для определения кусочно-непрерывной функции на каждом отрезке $[x_i, x_{i+1}]$ $(i = 0,1, ..., n-1)$ вводится такая локальная система координат, что $x_i = 0$ и $x_{i+1} = 1$. Тогда на каждом отрезке $[x_i, x_{i+1}]$ коэффициенты интерполирующего полинома

$$f(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 \quad (8.75)$$

можно найти из условий его прохождения на концах отрезка:

$$\left.\begin{aligned} f(0) &= a_0 \\ f(1) &= a_0 + a_1 + a_2 + a_3 \\ f'(0) &= a_1 \\ f'(1) &= a_1 + 2a_2 + 3a_3 \end{aligned}\right\}. \quad (8.76)$$

Отсюда следуют выражения для вычисления значений коэффициентов полинома:

$$\left.\begin{aligned} a_0 &= f(0) \\ a_1 &= f'(0) \\ a_2 &= 3[f(1) - f(0)] - 2f'(0) - f'(1) \\ a_3 &= 2[f(0) - f(1)] + f'(0) + f'(1) \end{aligned}\right\}. \qquad (8.77)$$

Значения первых двух коэффициентов определяются непосредственно, значения последних двух – из второго и третьего уравнений системы (8.76). Последовательно переходя от одного отрезка к другому, можно построить интерполирующую функцию на всем отрезке [a, b] (рис. 8.51), которая и будет локальным кубическим сплайном – кусочно-непрерывной функцией, имеющей на каждом отрезке $[x_i, x_{i+1}]$ вид

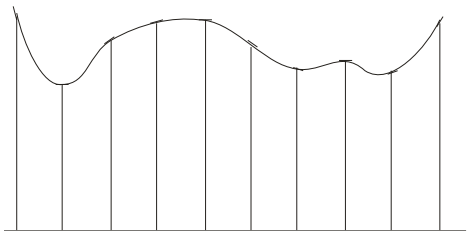


Рис. 8.51. Локальный сплайн

$$\sigma_i(x) = a_{i0} + a_{i1}x + a_{i2}x^2 + a_{i3}x^3. \qquad (8.78)$$

Простейшими являются линейные сплайны

$$\sigma_i(x) = a_{i0} + a_{i1}x, \qquad (8.79)$$

когда интерполирующая функция представляет собой ломаную.

Но при использовании кубических сплайнов остается проблема определения значений производной в узлах. Как правило, в исходных данных они отсутствуют (по крайней мере – при геомоделировании) и вычисляются тем или иным способом. Здесь, как это часто бывает в технике, одни проблемы заменяются другими. По поводу проблем, возникающих при интерполяции функций, П. Безье не без горечи, но, вероятно, с полным основанием замечал, что как бы тщательно ни был разработан алгоритм, всегда найдется частный случай, когда он не сработает. В подтверждение этого замечания он приводил пример, когда в точке P_i направление касательной к кривой принимается равным направлению хорды, соединяющей точки P_{i-1} и P_{i+1}.

Это, казалось бы, вполне разумное предположение может не сработать, когда кривизна кривой на участке меняет знак, что и было продемонстрировано П. Безье (рис. 8.52). В точке P_i направление касательной совпадает с направлением хорды $[P_{i-1}, P_{i+1}]$. Но некоторые программы могут построить кривую именно таким образом; это определяется их особенностями. Ошибка интерполирования, приведенная на рис. 8.52, может быть устранена путем подбора направления касательной в точке P_i (рис. 8.53). Но она также может быть исправлена путем подбора касательной в точке P_{i-1}.

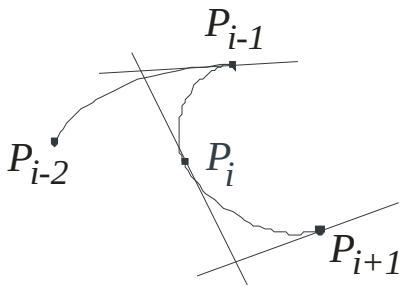


Рис. 8.52. Ошибка интерполирования

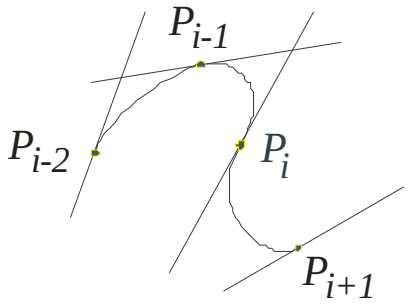


Рис. 8.53. Использование касательных

Результаты экспериментов показали, что в одних приложениях более высокую точность дают полные сплайны, в других – локальные. Последние представляют больший интерес, так как обеспечивают возможность манипулирования положением кривой в некоторой ограниченной области. В работе [16] указывалось, что полученный профиль земной поверхности может сильно отличаться от действительного положения, если в одной из точек производная имеет большое по модулю значение (см. рис. 8.49), либо расстояния между соседними узлами сильно различаются, что почти одно и то же. Для погашения ошибки интерполирования в подобных ситуациях предлагалось определять линейной интерполяцией дополнительные точки и кубический сплайн строить по совокупности исходных и дополнительных точек.

По этому поводу П. Безье отмечал, что данную проблему можно решить, задавая в точке x_i значения самой функции и ее первой и второй производных, а в качестве четвертого условия принимать значение функции в точке x_{i+1}, так как с математической точки зрения оба способа эквивалентны. Но он же признавал, что практически управлять формой кривой с помощью второй производной труднее.

8.16. Условия гладкости кусочно-непрерывных функций одной переменной

Несмотря на большое число имеющихся методов восполнения функций одной переменной, поиски эффективных методов для тех или иных конкретных условий продолжаются. Лучшие результаты получаются при использовании кусочно-непрерывных методов восполнения. Поэтому необходимо рассмотреть общие условия гладкости кусочно-непрерывных функций.

Определение таких условий начнем с функций одной переменной. Пусть на отрезке $[a, b]$ задана сетка

$$\Delta : a = x_0 < x_1 < \ldots < x_n = b, \qquad (8.80)$$

разбивающая его на n линейных конечных элементов

$$\{L_i : x_i \le x \le x_{i+1}\}\ (i = 0,1, \ldots, n-1). \qquad (8.81)$$

В узлах сетки задана своими значениями $h(x_i)$ $(i = 0,1, \ldots, n)$ приближаемая функция $h(x)$ – однозначная, непрерывная и гладкая (рис. 8.54).

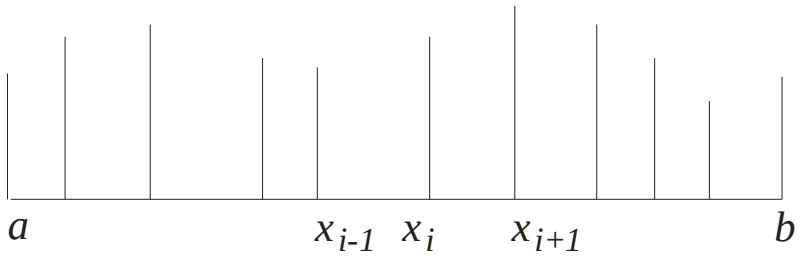


Рис. 8.54. Кусочно-непрерывная функция на сетке Δ

Требуется, чтобы приближающая функция $F(x)$ проходила точно через заданные значения функции $h(x)$:

$$F(x_i) = h(x_i) \ \ (i = 0,1, ..., n). \qquad (8.82)$$

Приближающую функцию $F(x)$ будем отыскивать в виде объединения кусочно-непрерывных функций

$$F(x) = \{F_i(x)\} \ \ (i = 0,1, ..., n-1) \qquad (8.83)$$

при условии, что областью определения функции $F_i(x)$ является соответствующий конечный элемент L_i.

В качестве исходной посылки примем, что интерполирующая функция $F_i(x)$ на элементе L_i имеет вид

$$F_i(x) = \varphi_i(x)h(x_i) + \varphi_{i+1}(x)h(x_{i+1}), \qquad (8.84)$$

где функции $\varphi(x)$ определяют вид приближающей функции и называются *функциями формы*, или *интерполяционными функциями* (рис. 8.55). Последнее выражение означает, что интерполяционная функция $F_i(x)$ на отрезке L_i является линейной комбинацией функций $\varphi(x)$, определенных на его концах.

Пусть известно, что приближаемая функция $h(x)$ в локальной области, содержащей точку x_i, с достаточной степенью точности может быть представлена некоторой имеющей непрерывную производную функцией $f_i(x)$ такой, что

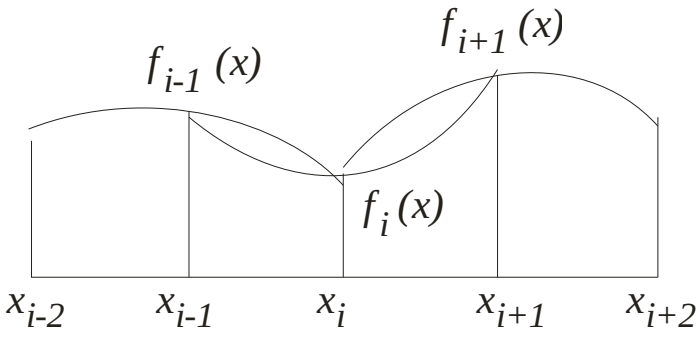


Рис. 8.55. Локальные функции

$$f_i(x_i) = h(x_i) \ , \qquad (8.85)$$

и точность представления ухудшается по мере удаления от x_i. Каждую функцию $f_i(x)$ будем называть *локальной*. Естественно предположить, что при движении точки *x* на отрезке $[x_i, x_{i+1}]$ от x_i к x_{i+1} «вклад» функции $f_i(x)$ в функцию $F_i(x)$ должен убывать, а функции $f_{i+1}(x)$ – возрастать, и наоборот. Поэтому целесообразно с каждой локальной функцией $f_i(x)$ связать понятие веса, характеризующего

степень доверия к ней. Кроме того, если $f_i(x)$ хорошо приближает функцию $h(x)$ в точке x_i, а $f_{i+1}(x)$ – в точке x_{i+1}, то мы имеем основания считать, что на элементе L_i функция $f_i(x)$ некоторым непрерывным образом *трансформируется* в $f_{i+1}(x)$.

Тогда интерполирующую функцию $F_i(x)$ определим как средневесовое двух смежных локальных функций $f_i(x)$ и $f_{i+1}(x)$:

$$F_i(x) = p_i(x)\, f_{\text{i}}(x) + p_{i+1}(x)\, f_{i+1}(x), \qquad (8.86)$$

где весовые функции $p_i(x)$ с областью определения L_i являются непрерывными, гладкими и отвечают условию

$$p_i(x) + p_{i+1}(x) = 1. \qquad (8.87)$$

Выражение (8.86) представляет собой более общий случай по сравнению с (8.84); в частности, если принять

$$\left.\begin{aligned} f_i(x) &= h(x_i) \\ f_{i+1}(x) &= h(x_{i+1}) \end{aligned}\right\}, \qquad (8.88)$$

то получим (8.84), при этом весовые функции будут играть роль функций формы.

Очевидно, что $F(x)$ непрерывна и имеет непрерывную первую производную во всех точках $x \in (x_i, x_{i+1})$, если на этом промежутке локальные и весовые функции являются гладкими. Чтобы $F(x)$ была непрерывной на всем отрезке $[a, b]$, необходимо, чтобы она была непрерывной и в точках стыковки x_i $(i = 1,2, ..., n-1)$ соседних элементов, для чего достаточно выполнения на каждом элементе L_i условий

$$\left.\begin{aligned} p_i(x_i)\, f_{_i}(x_i) + p_{i+1}(x_i)\, f_{i+1}(x_i) &= f_i(x_i) \\ p_i(x_{i+1})\, f_i(x_{i+1}) + p_{i+1}(x_{i+1})\, f_{i+1}(x_{i+1}) &= f_{i+1}(x_{i+1}) \end{aligned}\right\}. \qquad (8.89)$$

Последние равенства можно записать иначе:

$$\left.\begin{aligned} [\,p_i(x_i) - 1]\, f_{_i}(x_i) + p_{i+1}(x_i)\, f_{i+1}(x_i) &= 0 \\ p_i(x_{i+1})\, f_i(x_{i+1}) + [\,p_{i+1}(x_{i+1}) - 1]\, f_{i+1}(x_{i+1}) &= 0 \end{aligned}\right\}. \qquad (8.90)$$

Отсюда, с учетом (8.87), получаем

$$\left.\begin{aligned} p_{i+1}(x_i)[\,f_{i+1}(x_i) - f_i(x_i)] &= 0 \\ p_i(x_{i+1})[\,f_i(x_{i+1}) - f_{i+1}(x_{i+1})] &= 0 \end{aligned}\right\}. \qquad (8.91)$$

Из равенства (8.91) вытекают два различных способа получения непрерывной приближающей функции $F(x)$:

1) либо на концах конечного элемента L_i локальные функции принимают значения $f_i(x_i) = h(x_i)$ и $f_i(x_{i+1}) = h(x_{i+1})$, а весовые функции отвечают условиям

$p_i(x_j) = 1$ при $i = j$ и $p_i(x_j) = 0$ при $i \neq j$; (8.92)

2) либо каждая локальная функция $f_i(x)$ должна удовлетворять, помимо (8.85), условиям

$$\left.\begin{aligned} f_i(x_{i-1}) &= f_{i-1}(x_{i-1}) \\ f_i(x_{i+1}) &= f_{i+1}(x_{i+1}) \end{aligned}\right\}, \qquad (8.93)$$

то есть должно выполняться равенство значений двух соседних локальных функций в общем узле, а между весовыми функциями выполняется только соотношение (8.87).

В качестве условий непрерывности приближающей функции $F(x)$ в узлах сетки (8.80) примем равенства (8.92). Это означает, что каждая локальная функция $f_i(x)$ совпадает с приближаемой только в одном узле $f_i(x_i)$, а в остальных узлах это условие может не выполняться.

Так как на элементе L_i поведение интерполирующей функции зависит только от двух локальных функций на его концах, веса остальных локальных функций на этом элементе равны нулю, то есть веса являются финитными функциями. Тогда приближающую функцию $F(x)$ формально можно записать в виде

$$F(x) = \sum_{i=0}^{n} p_i(x)\, f_i(x). \qquad (8.94)$$

Данный результат можно получить другим путем. Известно, что интерполяционный многочлен имеет вид:

$$F(x) = \sum_{i=0}^{n} p_i(x) h(x_i), \qquad (8.95)$$

где $p_i(x)$ – такие многочлены степени *n*, что

$$p_i(x_j) = \begin{cases} 1 \text{ при } i = j \\ 0 \text{ при } i \neq j \end{cases}, \qquad (8.96)$$

а $h(x_i)$ – значения приближаемой функции.

Обычно функции $p(x)$ рассматриваются как полиномы постольку, поскольку речь идет об интерполировании полиномами, но это могут быть любые непрерывные функции, для которых выполняются соотношения (8.87). Таким образом, хотя функции $p_i(x)$ в точках x_j $(i \neq j)$ имеют нулевые значения, их поведение на отрезке $[x_j, x_{j+1}]$ в общем случае произвольно. Суммирование *n* произведений функций $p_i(x)$ на константы $h(x_i)$ часто дает на $[x_j, x_{j+1}]$ неожиданный и нежелательный результат, например, осцилляции. Положение интерполирующей функции будет более определенным и

контролируемым, если принять $p_i(x)=0$ для всех $x\in[x_j,x_{j+1}]$ $(i\neq j$ и $i\neq j+1)$. В итоге мы приходим к интерполированию кусочно-непрерывными функциями. Кроме того, если в (8.95) заменить константы $h(x_i)$ функциями $f_i(x)$, отвечающими условиям (8.85), то будет получено выражение (8.94).

На основании сравнения интерполирующих функций (8.94) и (8.95) может показаться, что первая из них в некоторой мере более ограничена по своим возможностям, так как поведение $p_i(x)$ на $[x_j,x_{j+1}]$ жестко регламентировано. Но в (8.95) функции $p_i(x)$ также фиксированы (их параметры определяются по заданным значениям в узлах сетки); разница лишь в том, что их поведение заранее не предсказуемо, что, скорее всего, следует считать нежелательным. Использование же (8.94) позволяет, подбирая локальные функции $f_i(x)$, добиваться нужного эффекта.

Преимущество функции (8.95) перед (8.94) – непрерывность первых n производных – на практике обычно не имеет решающего значения. Тем не менее, приближающая функция $F(x)$ иногда должна обладать некоторыми непрерывными производными, поэтому ниже рассматриваются условия их существования.

Гладкость приближающей функции в точке стыковки двух конечных элементов L_{i-1} и L_i означает равенство левого и правого пределов производной $F'(x)$ в узле x_i:

$$\lim_{(x-x_i)\to-0}F'_{i-1}(x)=\lim_{(x-x_i)\to+0}F'_i(x)=\lim F'(x_i) \quad . \qquad (8.97)$$

Продифференцировав (8.86), получим

$$F'_i(x)=p'_i(x)f_i(x)+p_i(x)f'_i(x)+$$

$$+p'_{i+1}(x)f_{i+1}(x)+p_{i+1}(x)f'_{i+1}(x). \qquad (8.98)$$

Так как $p_i(x_j)=0$ при $i\neq j,$ то значения левого и правого пределов производной $F'(x)$ в точке x_i будут вычисляться соответственно по формулам

$$\left.\begin{aligned}F'_{i-1}(x_i)&=p'_{i-1}(x_i)f_{i-1}(x_i)+p'_i(x_i)f_i(x_i)+f'_i(x_i)\\F'_i(x_i)&=p'_i(x_i)f_i(x_i)+f'_i(x_i)+p'_{i+1}(x_i)f_{i+1}(x_i)\end{aligned}\right\}. \qquad (8.99)$$

Следует обратить внимание на то, что в этих соотношениях $p'_i(x_i)$ по понятным причинам означают различные величины (для L_{i-1} и L_i соответственно).

Так как по исходным предположениям $f_i(x)$ достаточно хорошо приближает функцию $h(x)$ в окрестности точки x_i, естественно принять

$$F'(x_i) = F'_{i-1}(x_i) = F'_i(x_i) = f'_i(x_i). \qquad (8.100)$$

Из (8.100) и двух предшествующих выражений следует, что для обеспечения непрерывности $F'(x)$ в узле x_i должны выполняться соотношения

$$\left.\begin{aligned} p'_{i-1}(x_i)\, f_{i-1}(x_i) + p'_i(x_i)\, f_i(x_i) = 0 \\ p'_i(x_i)\, f_i(x_i) + p'_{i+1}(x_i)\, f_{i+1}(x_i) = 0 \end{aligned}\right\}. \qquad (8.101)$$

Для левого и правого конечных элементов из (8.87) следуют равенства

$$\left.\begin{aligned} p'_{i-1}(x) = -p'_i(x) \\ p'_i(x) = -p'_{i+1}(x) \end{aligned}\right\}, \qquad (8.102)$$

где опять-таки $p_i(x)$ являются разными функциями для L_{i-1} и L_i. Тогда на основании (8.101) и (8.102) получаем

$$\left.\begin{aligned} p'_{i-1}(x_i)[\, f_{i-1}(x_i) - f_i(x_i)] = 0 \\ p'_{i+1}(x_i)[\, f_{i+1}(x_i) - f_i(x_i)] = 0 \end{aligned}\right\}. \qquad (8.103)$$

Отсюда следуют два способа получения непрерывной в точке x_i производной $F'(x)$:

1) либо весовые функции $p_{i-1}(x)$ и $p_{i+1}(x)$ должны быть такими, что их производные при $x = x_i$ равны нулю;

2) либо локальные функции $f_{i-1}(x)$ и $f_{i+1}(x)$ в этой точке должны принимать значения

$$f_{i-1}(x_i) = f_{i+1}(x_i) = h(x_i). \qquad (8.104)$$

Таким образом, чтобы производная $F'(x)$ была непрерывна на всем отрезке $[a, b]$, необходимо, чтобы

$$\left.\begin{aligned} p'_i(x_{i-1}) = 0 \\ p'_i(x_{i+1}) = 0 \end{aligned}\right\}, \qquad (8.105)$$

или

$$\left.\begin{aligned} f_i(x_{i-1}) = f_{i-1}(x_{i-1}) \\ f_i(x_{i+1}) = f_{i+1}(x_{i+1}) \end{aligned}\right\}. \qquad (8.106)$$

Равенства (8.106) совпадают с (8.93). Ранее в качестве условий непрерывности мы выбрали соотношения (8.92). Выполнение равенств (8.93) или (8.106) обеспечивает непрерывность первой производной, поэтому будем считать их *условиями гладкости*. Следует обратить внимание, что при определении условий гладкости функции мы исходили из предположения, что выполняются соотношения (8.92).

Если локальные функции $f_i(x)$ и весовые функции $p_i(x)$ имеют непрерывную вторую производную, то интерполирующая функция будет обладать этим свойством при всех $x \in (x_i, x_{i+1})$. Непрерывность $F''(x)$ в точке стыковки x_i двух элементов L_{i-1} и L_i означает

$$\lim_{(x-x_i)\to -0} F''_{i-1}(x) = \lim_{(x-x_i)\to +0} F''_i(x) = \lim F''(x_i). \qquad (8.107)$$

Продифференцировав (8.98), получим

$$F''_i(x) = p''_i(x) f_i(x) + 2p'_i(x) f'_i(x) + p_i(x) f''_i(x) +$$

$$+ p''_{i+1}(x) f_{i+1}(x) + 2p'_{i+1}(x) f'_{i+1}(x) + p_{i+1}(x) f''_{i+1}(x). \qquad (8.108)$$

Тогда, с учетом (8.92) и (8.106), левый и правый пределы $F''(x)$ в узле x_i будут соответственно равны

$$\left.\begin{aligned} F''_{i-1}(x_i) &= p''_{i-1}(x_i) f_{i-1}(x_i) + 2p'_{i-1}(x_i) f'_{i-1}(x_i) + \\ &+ p''_i(x_i) f_i(x_i) + 2p'_i(x_i) f'_i(x_i) + f''_i(x_i) \\ F''_i(x_i) &= p''_i(x_i) f_i(x_i) + 2p'_i(x_i) f'_i(x_i) + f''_i(x_i) + \\ &+ p''_{i+1}(x_i) f_{i+1}(x_i) + 2p'_{i+1}(x_i) f'_{i+1}(x_i) \end{aligned}\right\}. \qquad (8.109)$$

Из (8.102) следует, что

$$\left.\begin{aligned} p''_{i-1}(x) &= -p''_i(x) \\ p''_i(x) &= -p''_{i+1}(x) \end{aligned}\right\}, \qquad (8.110)$$

где $p''_i(x)$ – также различные функции для L_{i-1} и L_i. На основании (8.102), (8.106) и (8.110) из выражений (8.109) получаем

$$\left.\begin{aligned} F''_{i-1}(x_i) &= 2p'_{i-1}(x_i)[f'_{i-1}(x_i) - f'_i(x_i)] + f''_i(x_i) \\ F''_i(x_i) &= 2p'_{i+1}(x_i)[f'_{i+1}(x_i) - f'_i(x_i)] + f''_i(x_i) \end{aligned}\right\}. \qquad (8.111)$$

Так как по начальным условиям функция $f_i(x)$ хорошо приближает $h(x)$ в окрестности точки x_i, примем, что

$$\left.\begin{aligned} F''_{i-1}(x_i) &= f''_i(x_i) \\ F''_i(x) &= f''_i(x_i) \end{aligned}\right\}. \qquad (8.112)$$

Подставив (8.112) в (8.111), находим, что интерполяционная функция $F(x)$ будет иметь непрерывную вторую производную в точке x_i, если

$$\left.\begin{aligned} p'_{i-1}(x_i)[f'_{i-1}(x_i) - f'_i(x_i)] &= 0 \\ p'_{i+1}(x_i)[f'_{i+1}(x_i) - f'_i(x_i)] &= 0 \end{aligned}\right\}. \qquad (8.113)$$

Следовательно, $F''(x)$ будет непрерывна на всем отрезке $[a,b]$, если во всех внутренних узлах выполняются требования:

1) либо

$$\left.\begin{array}{l} p'_i(x_{i-1}) = 0 \\ p'_i(x_{i+1}) = 0 \end{array}\right\}; \qquad (8.114)$$

2) либо

$$\left.\begin{array}{l} f'_i(x_{i-1}) = f'_{i-1}(x_{i-1}) \\ f'_i(x_{i+1}) = f'_{i+1}(x_{i+1}) \end{array}\right\}. \qquad (8.115)$$

Предпосылки большей гладкости не рассматриваются, так как способ их получения очевиден. Кроме того, для большинства приложений достаточно непрерывности интерполирующей функции и ее первой производной. Задача построения интерполирующей функции с непрерывной второй производной возникает сравнительно редко, например, при проектировании автомобильных и железных дорог горизонтальные и вертикальные кривые должны иметь непрерывную вторую производную, поскольку связаны с ускорением, которое должно быть непрерывным.

При выводе условий непрерывности $F''(x)$ в узлах сетки предполагалось выполнение равенств (8.92) и (8.106). В результате были получены соотношения (8.114), совпадающие с (8.105). Таким образом, требования к весовым функциям $p(x)$ и локальным функциям $f(x)$ в определенном смысле являются эквивалентными и дополняющими друг друга. Взаимосвязь между ними с одной стороны и свойствами гладкости интерполирующей функции $F(x)$ в узлах сетки Δ с другой представлена в табл. 8.5, которая может быть продолжена как вправо, путем наложения дополнительных ограничений на локальные функции $f(x)$, так и вниз, путем предъявления дополнительных требований к $p(x)$. В табл. 8.5 предполагается, что если выполняются условия A2 или B2, то выполняются и предшествующие им A1 или B1 и т. д.

Вопрос о выборе конкретных функций $p(x)$ и $f(x)$ может и должен решаться с учетом характера приближаемой функции и особенностей ее поведения на ограниченном участке или в точке. Очевидно, что локальные функции в двух смежных узлах могут быть различного вида, но с целью упрощения они могут иметь и один тип. Равным образом могут быть различными весовые функции на двух смежных элементах. Легко видеть, что $f_i(x)$ и $f_{i+1}(x)$ являются верхней и нижней границами для результирующей функции $F_i(x)$. Положение частной функции $F_i(x)$ по отношению к локальным функциям $f_i(x)$ и $f_{i+1}(x)$ и, следовательно, окончательное положение $F(x)$ на элементе L_i определяется видом весовых функций.

Таблица 8.5. Условия непрерывности кусочно-непрерывной функции и ее производных

Свойства весовых функций	Свойства локальных функций		
	$f_i(x_i) = h(x_i)$ (B1)	$f_i(x_i) = f_{i+1}(x_i)$ $f_i(x_{i+1}) = f_{i+1}(x_{i+1})$ (B2)	$f'_i(x_i) = f'_{i+1}(x_i)$ $f'_i(x_{i+1}) = f'_{i+1}(x_{i+1})$ (B3)
$p_i(x) + p_{i+1}(x) = 1$ (A1)		$F(x)$	$F'(x)$
$p_i(x_j) = \begin{cases} 1 (i = j) \\ 0 (i \neq j) \end{cases}$ (A2)	$F(x)$	$F'(x)$	$F''(x)$
$p'_i(x_i) = 0$ $p'_{i+1}(x_i) = 0$ $p'_i(x_{i+1}) = 0$ $p'_{i+1}(x_{i+1}) = 0$ (A3)	$F'(x)$	$F''(x)$	$F'''(x)$

Примерами весовых функций могут служить следующие пары:

$$\left.\begin{aligned} p_i(x) &= \frac{x_{i+1} - x}{x_{i+1} - x_i} \\ p_{i+1}(x) &= \frac{x - x_i}{x_{i+1} - x_i} \end{aligned}\right\}; \qquad (8.116)$$

$$\left.\begin{aligned} p_i(x) &= \frac{(x_{i+1} - x)^2}{(x_{i+1} - x)^2 + (x - x_i)^2} \\ p_{i+1}(x) &= \frac{(x - x_i)^2}{(x_{i+1} - x)^2 + (x - x_i)^2} \end{aligned}\right\}; \qquad (8.117)$$

$$\left.\begin{aligned} p_i(x) &= \cos^2 \frac{\pi(x - x_i)}{2(x_{i+1} - x_i)} \\ p_{i+1}(x) &= \sin^2 \frac{\pi(x - x_i)}{2(x_{i+1} - x_i)} \end{aligned}\right\}; \qquad (8.118)$$

$$\left.\begin{aligned} p_i(x) &= 1 - 3\left(\frac{x - x_i}{x_{i+1} - x_i}\right)^2 + 2\left(\frac{x - x_i}{x_{i+1} - x_i}\right)^3 \\ p_{i+1}(x) &= 3\left(\frac{x - x_i}{x_{i+1} - x_i}\right)^2 - 2\left(\frac{x - x_i}{x_{i+1} - x_i}\right)^3 \end{aligned}\right\} \qquad (8.119)$$

– и ряд других.

Выбирая вид весовых и локальных функций, можно получить приближающую функцию с определенными свойствами. Ниже демонстрируются возможности описанного подхода на нескольких простых примерах.

1. Веса изменяются по линейному закону (8.116). Локальные функции $f_i(x)$ и $f_{i+1}(x)$ являются окружностями, проходящими через три соседние точки. (Такие попытки известны, но это не значит, что их следует повторять. Данный метод допустим только при частом расположении точек, когда расстояние между ними намного меньше радиуса кривизны.)

2. Весовые функции являются линейными, а локальные функции – константами $f_i(x) = h(x_i)$ и $f_{i+1}(x) = h(x_{i+1})$. Тогда интерполирующая функция $F(x)$ будет линейным сплайном.

3. Весовые функции линейны, локальные функции являются полиномами второго порядка

$$\left.\begin{aligned} f_i(x) &= a_i + b_i(x - x_i) + c_i(x - x_i)^2 \\ f_{i+1}(x) &= a_{i+1} + b_{i+1}(x - x_{i+1}) + c_{i+1}(x - x_{i+1})^2 \end{aligned}\right\}, \qquad (8.120)$$

проходящими через три соседние точки интерполируемой функции. Интерполирующая функция $F(x)$ будет являться локальным кубическим сплайном.

4. Узел x_i является точкой локального экстремума. Функцию $f_i(x)$ можно представить как кубический полином, проходящий через три смежные точки и имеющий в точке x_i первую производную, равную нулю.

Возможно другое решение: принимается $f_i(x) = h(x_i)$, а на элементах L_{i-1} и L_i в качестве весовых принимаются функции (8.118) или (8.119). В результате такого приближения точка локального экстремума $F(x)$ будет совпадать с узлом x_i .

6. Узел x_i является точкой разрыва первой производной. Тогда в качестве $f_i(x)$ может быть выбрана обладающая таким же свойством подходящая функция, а весовыми могут быть функции (8.116).

7. Интерполируемая функция задана в узлах x_i своими значениями $h(x_i)$ и значениями первой производной $h'(x_i)$. Пусть требуется, чтобы интерполирующая функция также имела непрерывную производную. Тогда локальные функции можно определить как кубические полиномы, проходящие через три соседние точки и имеющие в среднем узле заданное значение производной. Веса могут вычисляться по формуле (8.116).

8. Приближаемая функция представлена значениями $h(x_i)$ в узлах сетки, приближающая функция должна иметь непрерывную вторую производную. Локальные функции можно определить как параболы,

проходящие через три точки, а весовые функции – как (8.117), (8.118) или (8.119).

9. Если в узлах сетки моделируемая функция задана своими значениями и значениями первой производной, а вторая производная приближающей функции $F(x)$ должна быть непрерывна, то возможны два наиболее простых решения. В первом случае локальные функции представляются полиномами третьего порядка, коэффициенты которых определяются из условий

$$\left.\begin{aligned} &a_0 + a_1 x_{i-1} + a_2 x_{i-1}^2 + a_3 x_{i-1}^3 = h(x_{i-1}) \\ &a_0 + a_1 x_i + a_2 x_i^2 + a_3 x_i^3 = h(x_i) \\ &a_0 + a_1 x_{i+1} + a_2 x_{i+1}^2 + a_3 x_{i+1}^3 = h(x_{i+1}) \\ &a_1 + 2a_2 x_i + 3a_3 x_i^2 = h'(x_i) \end{aligned}\right\}, \qquad (8.121)$$

а для вычисления весов используются любые из выражений (8.117)–(8.119) или эквивалентные им.

Во втором случае локальные функции $f_i(x)$ определяются как полиномы пятой степени, коэффициенты которых находятся из условий

$$\left.\begin{aligned} &\sum_{j=0}^{5} a_j x_{i-1}^j = h(x_{i-1}) \\ &\sum_{j=0}^{5} a_j x_i^j = h(x_i) \\ &\sum_{j=0}^{5} a_j x_{i+1}^j = h(x_{i+1}) \\ &\sum_{j=1}^{5} j a_j x_{i-1}^{j-1} = h'(x_{i-1}) \\ &\sum_{j=1}^{5} j a_j x_i^{j-1} = h'(x_i) \\ &\sum_{j=1}^{5} j a_j x_{i+1}^{j-1} = h'(x_{i+1}) \end{aligned}\right\}, \qquad (8.122)$$

а веса вычисляются по формулам (8.116).

Использование трех соседних точек для построения гладкой локальной функции $f_i(x)$ не является обязательным, $f_i(x)$ может строиться и по большему числу точек. Но в некоторых приложениях, когда поведение $h(x)$ не является монотонным, это решение может оказаться принципиальным.

Возможно также получение интерполирующей функции как среднего весового не двух, а большего числа локальных функций, то есть

$$F(x) = \sum_{i=1}^{k} p_i(x)\, f_i(x), \qquad (8.123)$$

где $k > 2$. Но развитие рассмотренного метода интерполяции в этом направлении едва ли целесообразно, так как сопровождается значительным возрастанием его сложности при одновременном снижении эффективности.

Некоторые особенности возникают при обработке первого и последнего конечных элементов. Но здесь также предоставляются различные возможности. Так, например, $f_0(x)$ может определяться по двум точкам или совпадать с функцией $f_1(x)$ и т. п.

В многочисленных имеющихся методах восполнения функция одной переменной на всех конечных элементах представляется функцией одного и того же вида. Описанный подход снимает это ограничение. Любые две локальные функции $f_i(x)$ и $f_{i+1}(x)$ могут быть различного типа. Например, одна из них может быть полиномом некоторой степени, а другая – спиралью или рациональной функцией и т. д. Данный метод, видимо, является единственным, обеспечивающим постепенную трансформацию функции одного вида в функцию другого вида.

Метод может применяться для интерполирования кривых. Возникающие при этом затруднения устраняются либо введением независимого параметра *t*

$$\left.\begin{aligned} x &= x(t) \\ y &= y(t) \end{aligned}\right\}, \qquad (8.124)$$

либо посредством перехода к локальной системе координат, либо разбиением кривой на участки, на каждом из которых она может быть представлена однозначной функцией. Использование независимого параметра представляется более практичным решением.

Пожалуй, наибольший интерес метод представляет из-за его алгоритмичности. Программа может быть организована таким образом, что блоки, вычисляющие параметры и значения весовых и локальных функций, вызываются по мере необходимости, сменяя друг друга. При желании пользователь может расширить возможности такой программы, написав собственные модули для вычисления $p(x)$ и $f(x)$ и подключив их к программе.

Быстродействие такой универсальной программы интерполирования будет ниже, чем быстродействие специализированных программ, реализующих единственный способ приближения функций. Но такая программа

будет беспрецедентной по своим возможностям и должна сыграть положительную роль в условиях недостаточного количества программных средств и их высокой стоимости. Кроме того, универсальная программа может служить в качестве испытательного стенда при отборе методов приближения функций одной переменной. Можно написать несколько модулей для локальных и весовых функций, выбрать по результатам экспериментов наилучшее их

сочетание и затем написать высокоэффективную специализированную программу.

Наконец, установленные условия непрерывности интерполирующей функции и ее производных избавляют от необходимости каждый раз при разработке нового метода кусочно-непрерывной интерполяции доказывать свойства получаемой функции. При реализации системы моделирования топографических поверхностей описанный подход может быть использован для восполнения горизонталей, получаемых отслеживанием по сетке прямоугольников или произвольных треугольников.

8.17. Интерполирование кривых

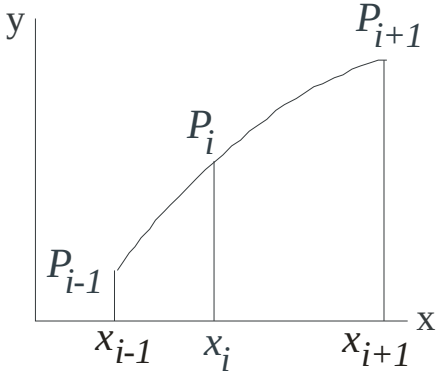


Рис. 8.57. Интерполяция параболой

Задача интерполирования кривых принципиально отличается от задачи интерполирования функций тем, что первые, в отличие от последних, являются неоднозначными функциями и могут быть как разомкнутыми, так и замкнутыми.

Пусть замкнутая интерполируемая кривая задана конечным набором своих точек $P_i = (x_i, y_i) \quad (i = 1, ..., n)$ (рис. 8.56). Первый способ интерполяции кривой состоит в том, чтобы свести эту задачу к интерполяции функции. Для этого кривая разбивается на участки, на каждом из которых она может рассматриваться как однозначная функция $y = f(x)$ либо как $x = f(y)$. Если на участке кривой $(P_j, ..., P_k)$ для каждой пары смежных точек выполняется одно и только одно из строгих неравенств $x_i < x_{i+1}$ или $x_i > x_{i+1}$ $(i = j, ..., k-1)$, то на этом участке кривая может быть представлена как однозначная функция $y = f(x)$. Если на участке кривой выполняются аналогичные условия для y, то есть либо $y_i < y_{i+1}$, либо $y_i > y_{i+1}$ $(i = j, ..., k-1)$, то кривая на нем может быть представлена как функция $x = f(y)$. Определив на концах каждого участка в качестве граничных условий направление касательной к кривой и/или другие условия, можно получить всюду гладкую замкнутую кривую.

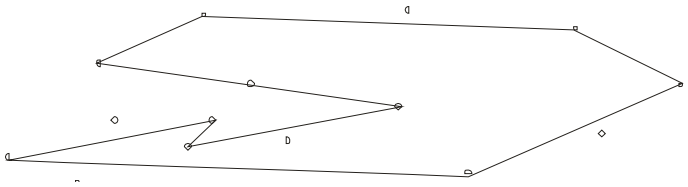


Рис. 8.56. Интерполирование кривой

Рассмотрение интерполирования кривых (впрочем, как и функций одной переменной) следует начинать с простейшего случая. Задача интерполирования функции или кривой начинается с добавления третьей точки, которая не лежит на прямой, проведенной через две заданные.

Параболическая интерполяция. Обычное решение задачи проведения гладкой кривой через три точки P_{i-1}, P_i и P_{i+1} (рис. 8.57) заключается в том, что через них проводится парабола

$$y = a_0 + a_1 x + a_2 x^2. \qquad (8.125)$$

Неизвестные коэффициенты этого уравнения могут быть получены из условия прохождения функции через три заданные точки

$$\left.\begin{aligned} a_0 + a_1 x_{i-1} + a_2 x_{i-1}^2 &= y_{i-1}, \\ a_0 + a_1 x_i + a_2 x_i^2 &= y_i, \\ a_0 + a_1 x_{i+1} + a_2 x_{i+1}^2 &= y_{i+1}. \end{aligned}\right\} \qquad (8.126)$$

Вычитая второе уравнение из первого и третьего, получим систему линейных уравнений с двумя неизвестными

$$a_1(x_{i-1} - x_i) + a_2(x_{i-1}^2 - x_i^2) = y_{i-1} - y_i;$$

$$a_1(x_{i+1} - x_i) + a_2(x_{i+1}^2 - x_i^2) = y_{i+1} - y_i.$$

Чтобы исключить a_1, каждое полученное равенство умножим слева и справа

$$a_1(x_{i-1} - x_i)(x_{i+1} - x_i) + a_2(x_{i-1}^2 - x_i^2)(x_{i+1} - x_i) = (y_{i-1} - y_i)(x_{i+1} - x_i);$$

$$a_1(x_{i-1} - x_i)(x_{i+1} - x_i) + a_2(x_{i-1} - x_i)(x_{i+1}^2 - x_i^2) = (y_{i+1} - y_i)(x_{i-1} - x_i)$$

и из второго равенства вычтем первое

$$a_2[(x_{i-1} - x_i)(x_{i+1}^2 - x_i^2) - (x_{i-1}^2 - x_i^2)(x_{i+1} - x_i)] =$$

$$= (y_{i+1} - y_i)(x_{i-1} - x_i) - (y_{i-1} - y_i)(x_{i+1} - x_i).$$

Отсюда находим значение коэффициента

$$a_2 = \frac{(y_{i+1} - y_i)(x_{i-1} - x_i) - (y_{i-1} - y_i)(x_{i+1} - x_i)}{(x_{i-1} - x_i)(x_{i+1}^2 - x_i^2) - (x_{i-1}^2 - x_i^2)(x_{i+1} - x_i)} \qquad (8.127)$$

или

$$a_2 = \frac{(y_{i+1} - y_i)(x_{i-1} - x_i) - (y_{i-1} - y_i)(x_{i+1} - x_i)}{(x_{i+1} - x_{i-1})(x_{i+1} - x_i)(x_{i-1} - x_i)}. \qquad (8.128)$$

Данное соотношение можно также представить как

$$a_2 = \frac{y_{i-1}(x_i - x_{i+1}) + y_i(x_{i+1} - x_{i-1}) + y_{i+1}(x_{i-1} - x_i)}{(x_{i+1} - x_{i-1})(x_{i+1} - x_i)(x_{i-1} - x_i)}.$$

Аналогичным образом можно определить значение коэффициента a_1:

$$a_1 = \frac{(y_{i+1} - y_i)(x_{i-1}^2 - x_i^2) - (y_{i-1} - y_i)(x_{i+1}^2 - x_i^2)}{(x_{i-1} - x_{i+1})(x_{i+1} - x_i)(x_{i-1} - x_i)} \qquad (8.129)$$

или

$$a_1 = \frac{y_{i-1}(x_i^2 - x_{i+1}^2) + y_i(x_{i+1}^2 - x_{i-1}^2) + y_{i+1}(x_{i-1}^2 - x_i^2)}{(x_{i-1} - x_{i+1})(x_{i+1} - x_i)(x_{i-1} - x_i)}.$$

Значения a_0 и a_1 можно также найти из уравнений (8.126) и (8.127)

$$a_1 = \frac{(y_{i+1} - y_i) - a_2(x_{i+1}^2 - x_{i-1}^2)}{x_{i+1} - x_{i-1}}; \qquad (8.130)$$

$$a_0 = \frac{y_{i-1}x_{i+1} - y_{i+1}x_{i-1} - a_2(x_{i-1}^2 x_{i+1} - x_{i-1}x_{i+1}^2)}{x_{i+1} - x_{i-1}}. \qquad (8.131)$$

Однако изложенный подход обладает определенным механицизмом, излишней прямолинейностью и неэффективностью. Поэтому интерполирование полиномами невысокой степени необходимо рассмотреть более подробно.

Первое, что можно сделать для повышения эффективности, – это перенести начало системы координат в точку P_i

$$\left.\begin{aligned} X_j &= x_j - x_i \\ Y_j &= y_j - y_i \end{aligned}\right\} \ (j = (i-1, i, i+1)).$$

В новой системе координат XY первое уравнение будет иметь вид

$$Y = a_0 + a_1 X + a_2 X^2. \qquad (8.132)$$

Очевидно, что $X_i = 0$ и $Y_i = 0$. Тогда коэффициенты уравнения можно получить из условий

$$\left.\begin{aligned} a_0 + a_1 X_{i-1} + a_2 X_{i-1}^2 &= Y_{i-1} \\ a_0 &= 0 \\ a_0 + a_1 X_{i+1} + a_2 X_{i+1}^2 &= Y_{i+1} \end{aligned}\right\}. \qquad (8.133)$$

Значения коэффициентов равны

$$\left.\begin{aligned} a_0 &= 0 \\ a_1 &= \frac{X_{i+1}^2 Y_{i-1} - X_{i-1}^2 Y_{i+1}}{(X_{i+1} - X_{i-1}) X_{i-1} X_{i+1}} \\ a_2 &= \frac{X_{i-1} Y_{i+1} - X_{i+1} Y_{i-1}}{(X_{i+1} - X_{i-1}) X_{i-1} X_{i+1}} \end{aligned}\right\}. \qquad (8.134)$$

В данном случае формулы для a_1 и a_2 можно получить из более общих формул (8.128) и (8.129), если учесть, что $X_i = 0$ и $Y_i = 0$.

Максимальное отклонение. В результате переноса начала системы координат удалось упростить выражения, но ничего нельзя сказать о качестве приближения, за исключением того, что полученная парабола проходит через три заданные точки. Чтобы внести некоторую ясность, рассмотрим следующий случай. Пусть интерполирующая функция, проходящая через точки P_{i-1}, P_i и P_{i+1}, является не только непрерывной и дифференцируемой во всех точках отрезка $[x_{i-1}, x_{i+1}]$, но и ее кривизна на этом отрезке сохраняет знак (рис. 8.58). В соответствии с теоремой Лагранжа на отрезке $[x_{i-1}, x_{i+1}]$ найдется точка, в которой производная параллельна хорде $[P_{i-1}, P_{i+1}]$. Пусть такой точкой является P_i. По условию, кривизна во всех точках дуги $[P_{i-1}, P_{i+1}]$ сохраняет свой знак, поэтому в любой точке дуги все остальные ее точки будут лежать по одну сторону от касательной к дуге в этой точке. Следовательно, точка P_i является точкой максимального отклонения δ дуги кривой от хорды $[P_{i-1}, P_{i+1}]$.

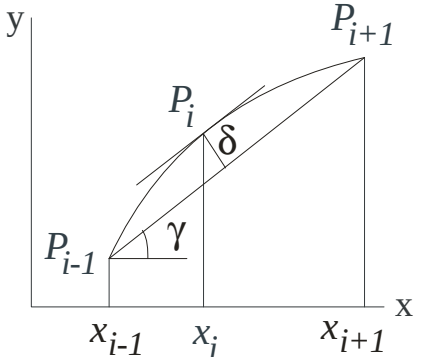


Рис. 8.58. Максимальное отклонение

Отклонение δ связано с величиной Δ (рис. 8.58) соотношением $\delta = \Delta \cos\gamma$, где

$$\Delta = f(x) - [y_{i-1} + \frac{y_{i+1} - y_{i-1}}{x_{i+1} - x_{i-1}}(x - x_{i-1})], \qquad (8.135)$$

а угол γ является константой и

$$\cos\gamma = \frac{x_{i+1} - x_{i-1}}{\sqrt{(x_{i+1} - x_{i-1})^2 + (y_{i+1} - y_{i-1})^2}}.$$

Отсюда следует, что для нахождения максимального отклонения δ достаточно найти максимальное значение Δ, которое может быть получено из условия

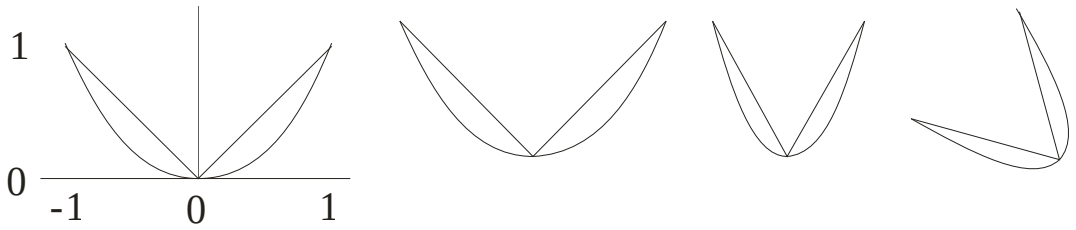
$$f'(x) - \frac{y_{i+1} - y_{i-1}}{x_{i+1} - x_{i-1}} = 0.$$

Для нас важно то, что если направление касательной в каждой точке P_i интерполируемой функции принимается равным направлению хорды (P_{i-1}, P_{i+1}), то такое решение можно признать достаточно разумным. Оно дает возможность ослабить эффекты, подобные изображенному на рис. 8.49. Однако при интерполяции параболой нет возможности управлять направлением

касательной в точке P_i , ее коэффициенты определяются однозначным образом из решения системы линейных уравнений (8.126).

Интерполирование на отрезке. Чтобы еще упростить задачу, рассмотрим случай, когда ломаная симметрична и задана координатами трех точек (-1,1), (0,0), (1,1) (рис. 8.59, а).

Первое, что можно сказать о представлении кривой, проходящей через заданные три точки, – это то, что в данном случае наиболее подходящей кривой является парабола $y = x^2$. Наши намерения не могут вызывать резких возражений. Функция $y = x^2$ симметрична и довольно проста. В точке (0,0) касательная к кривой параллельна хорде, соединяющей две другие точки. Кроме того, кривизна в этой точке имеет максимальное значение, что согласуется с изложенными выше соображениями о выборе точек на кривой, представляющих ее наилучшим образом. Точка (0,0) имеет максимальное отклонение от прямой, проведенной через первую и последнюю точки кривой.



а) б) в) г)

Рис. 8.59. Парабола y = x2

При линейных преобразованиях координат заданной тройки точек (смещении, сжатии и поворотах, что отражено на рис. 8.59, б, в, г), изменяя соответствующим образом систему координат, можно по-прежнему осуществлять интерполирование с помощью параболы. Во всех этих случаях ломаная остается симметричной, поскольку выполняется равенство двух ее отрезков. Вообще же, функция, проходящая через три точки P_{i-1}, P_i и P_{i+1}, имеющая в точке P_i максимальное отклонение от хорды (P_{i-1}, P_{i+1}) и экстремальное значение кривизны, будет «хорошей».

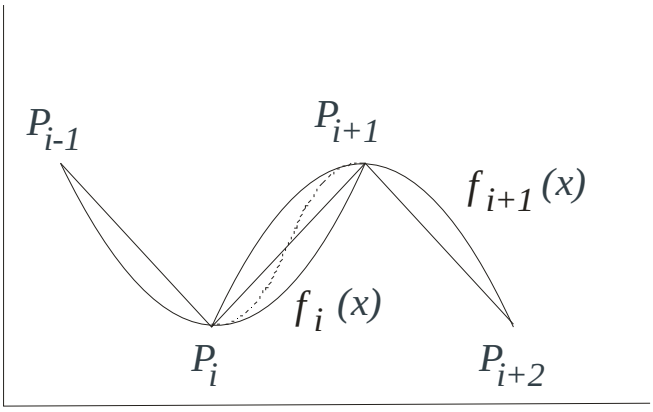


Рис. 8.60. Интерполирование параболами

Пусть требуется провести интерполирующую кривую через четыре точки (рис. 8.60). Допустим, что первая тройка точек лежит на параболе $f_i(x)$, а вторая тройка является смещенным зеркальным отображением первой, и ее

интерполирование может быть осуществлено параболой $f_{i+1}(x)$. Очевидно, что парабола $f_i(x)$ достаточно точно представляет интерполируемую кривую лишь вблизи точки P_i. По мере удаления от P_i качество интерполяции параболой $f_i(x)$ ухудшается. Хотя $f_i(x)$ проходит через точку P_{i+1}, нельзя считать, что она хорошо приближает интерполируемую кривую в этой точке. Качество приближения кривой в точке определяется не только степенью близости интерполирующей кривой к этой точке, но и направлением касательной к кривой.

То же самое можно сказать о восполнении кривой параболой $f_{i+1}(x)$, которая является хорошим приближением кривой в окрестности точки P_{i+1}, но при удалении от этой точки качество интерполирования снижается. Можно сделать вывод о том, что на отрезке $[P_i, P_{i+1}]$ интерполирующая кривая должна занимать некоторое промежуточное значение между $f_i(x)$ и $f_{i+1}(x)$ (см. рис. 8.60). В окрестности точки P_i интерполирующая кривая должна быть ближе к $f_i(x)$, а при движении к точке P_{i+1} она должна приближаться к $f_{i+1}(x)$. Таким образом, мы приходим к понятию веса и представлению интерполирующей функции $F_i(x)$ на отрезке $[P_i, P_{i+1}]$ как среднего весового функций $f_i(x)$ и $f_{i+1}(x)$. Собственно говоря, эти соображения были положены в основу рассмотренного выше способа получения гладких кусочно-непрерывных функций одной переменной, когда на отрезке $[x_i, x_{i+1}]$ функция $F_i(x)$ определялась как

$$F_i = p_i(x)\, f_i(x) + p_{i+1}(x)\, f_{i+1}(x), \qquad (8.136)$$

где $p(x)$ – весовые функции, а $f(x)$ – локальные функции.

Если на отрезке $[x_i, x_{i+1}]$ весовые функции определить как линейные

$$\left.\begin{aligned} p_i(x) &= \frac{x_{i+1} - x}{x_{i+1} - x_i} \\ p_{i+1}(x) &= \frac{x - x_i}{x_{i+1} - x_i} \end{aligned}\right\}, \qquad (8.137)$$

а локальные функции $f_i(x)$ – как параболы вида

$$f_i(x) = a_0^{(i)} + a_1^{(i)} x + a_2^{(i)} x^2, \qquad (8.138)$$

проходящие через точки (P_{i-1}), (P_i) и (P_{i+1}), то

$$p_i(x)\, f_i(x) = \frac{x_{i+1} - x}{x_{i+1} - x_i}\left(a_0^{(i)} + a_1^{(i)} x + a_2^{(i)} x^2\right);$$

$$p_{i+1}(x)\,f_{i+1}(x) = \frac{(x - x_i)}{(x_{i+1} - x_i)}(a_0^{(i+1)} + a_1^{(i+1)}x + a_2^{(i+1)}x^2),$$

и мы получаем один из вариантов локальных кубических сплайнов

$$F_i(x) = \frac{x_{i+1} - x}{x_{i+1} - x_i}(a_0^{(i)} + a_1^{(i)}x + a_2^{(i)}x^2) + \frac{x - x_i}{x_{i+1} - x_i}(a_0^{(i+1)} + a_1^{(i+1)}x + a_2^{(i+1)}x^2)$$

или

$$F_i(x) = \frac{1}{x_{i+1} - x_i}[(a_0^{(i)}x_{i+1} - a_0^{(i+1)}x_i) +$$

$$+ (a_1^{(i)}x_{i+1} - a_0^{(i)} + a_0^{(i+1)} - a_1^{(i+1)}x_i)x +$$

$$+ (a_2^{(i)}x_{i+1} - a_1^{(i)} + a_1^{(i+1)} - a_2^{(i+1)}x_i)x^2 + (a_2^{(i+1)} - a_2^{(i)})x^3]. \quad (8.139)$$

Приняв для констант обозначения

$$\left.\begin{aligned} b_0^{(i)} &= \frac{1}{x_{i+1} - x_i}(a_0^{(i)}x_{i+1} - a_0^{(i+1)}x_i) \\ b_1^{(i)} &= \frac{1}{x_{i+1} - x_i}(a_1^{(i)}x_{i+1} - a_0^{(i)} + a_0^{(i+1)} - a_1^{(i+1)}x_i) \\ b_2^{(i)} &= \frac{1}{x_{i+1} - x_i}(a_2^{(i)}x_{i+1} - a_1^{(i)} + a_1^{(i+1)} - a_2^{(i+1)}x_i) \\ b_3^{(i)} &= \frac{1}{x_{i+1} - x_i}(a_2^{(i+1)} - a_2^{(i)}) \end{aligned}\right\}, \quad (8.140)$$

интерполирующую функцию можно представить как

$$F_i(x) = b_0^{(i)} + b_1^{(i)}x + b_2^{(i)}x^2 + b_3^{(i)}x^3. \quad (8.141)$$

Полученные выражения можно упростить, если начало координат перенести в точку P_i. Тогда

$$F_i(x) = \frac{1}{x_{i+1}}[a_0^{(i)}x_{i+1} + (a_1^{(i)}x_{i+1} - a_0^{(i)} + a_0^{(i+1)})x +$$

$$+ (a_2^{(i)}x_{i+1} - a_1^{(i)} + a_1^{(i+1)})x^2 + (-a_2^{(i+1)} - a_2^{(i)})x^3]. \quad (8.142)$$

На отрезке $[P_i, P_{i+1}]$ коэффициенты функции $f_i(x) = a_0^{(i)} + a_1^{(i)}x + a_2^{(i)}x^2$ находят из условий (8.126) ее прохождения через три заданные точки.

Динамические функции. На интерполирование функций мы можем взглянуть с иной точки зрения. Можно считать, что функция $F_i(x)$ является «динамической», и в узле x_i она представляется как $f_i(x)$, а в узле x_{i+1} – как

$f_{i+1}(x)$. В промежутке между этими двумя узлами $F_i(x)$ меняется некоторым непрерывным образом. Так, например, функция

$$f_i(x) = a_0^{(i)} + a_1^{(i)} x + a_2^{(i)} x^2$$

трансформируется в функцию

$$f_{i+1}(x) = a_0^{(i+1)} + a_1^{(i+1)} x + a_2^{(i+1)} x^2 .$$

Если трансформируемая функция есть полином, то мы имеем дело с преобразованием полинома, а изменение полинома есть изменение его коэффициентов.

Мы можем предположить, что на отрезке $[x_i, x_{i+1}]$ коэффициенты полинома изменяются по линейному закону

$$b_j^i = a_j^i + \frac{x - x_i}{x_{i+1} - x_i}(a_j^{i+1} - a_j^i) \quad (j = 0, 1, 2, 3) \qquad (8.143)$$

при интерполировании вперед или

$$b_j^i = a_j^{i+1} - \frac{x - x_{i+1}}{x_i - x_{i+1}}(a_j^{i+1} - a_j^i) \quad (j = 0, 1, 2, 3) \qquad (8.144)$$

при интерполировании назад. Тогда локальная функция $F_i(x)$ на отрезке $[x_i, x_{i+1}]$ будет иметь вид

$$F_i(x) = \frac{1}{x_{i+1} - x_i}[b_0^{(i)} + b_1^{(i)} x + b_2^{(i)} x^2 + b_3^{(i)} x^3] , \qquad (8.145)$$

а если начало системы координат перенести в точку P_i , то при интерполировании вперед может использоваться формула

$$b_j^{(i)} = a_j^{(i)} + \frac{x}{x_{i+1}}(a_j^{(i+1)} - a_j^{(i)}) \quad (j = 0, 1, 2, 3),$$

а при интерполировании назад

$$b_j^{(i)} = a_j^{(i+1)} - \frac{x_{i+1} - x}{x_{i+1}}(a_j^{(i+1)} - a_j^{(i)}) \quad (j = 0, 1, 2, 3).$$

Следовательно, локальную функцию $F_i(x)$ на отрезке $[x_i, x_{i+1}]$ можно представить как

$$F_i(x) = \frac{1}{x_{i+1}}[b_0^{(i)} + b_1^{(i)} x + b_2^{(i)} x^2 + b_3^{(i)} x^3] . \qquad (8.146)$$

Линейная интерполяция полиномов. Понятие интерполирования значений функций может быть обобщено до понятия интерполирования функций. Частным случаем такого интерполирования является интерполирование полиномов. Если на концах отрезка $[x_i, x_{i+1}]$ определены полиномы n-й

степени $P_i^n(x)$ и $P_{i+1}^n(x)$, то в промежутке между этими узлами можно определить полином

$$P_i^{n+1}(x) = P_i^n(x) + \frac{P_{i+1}^n(x) - P_i^n(x)}{x_{i+1} - x_i}(x - x_i) \quad . \qquad (8.147)$$

Тогда линейная интерполяция может последовательно применяться для получения интерполирующих полиномов более высокой степени. Пусть, например, в узле x_i локальная функция имеет вид $f_i(x) = y_i$, а в узле x_{i+1} – вид $f_{i+1}(x) = y_{i+1}$ (рис. 8.61). Если теперь для получения интерполирующей функции применить интерполирование вперед

$$F_i(x) = f_i(x) + \frac{f_{i+1}(x) - f_i(x)}{x_{i+1} - x_i}(x - x_i) \quad , \qquad (8.148)$$

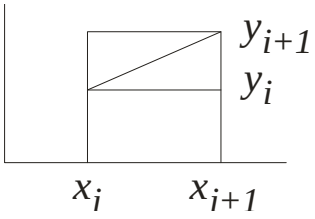


Рис. 8.61. Линейная интерполяция

то получим формулу обычной линейной интерполяции значений функции

$$F_i(x) = y_i + \frac{y_{i+1} - y_i}{x_{i+1} - x_i}(x - x_i) \quad .$$

Пусть функция задана своими значениями в узлах x_{i-1}, x_i и x_{i+1} (рис. 8.62). На отрезке $[x_{i-1}, x_i]$ определена локальная функция

$$f_{i-1}(x) = y_{i-1} + \frac{y_i - y_{i-1}}{x_i - x_{i-1}}(x - x_{i-1})$$

,

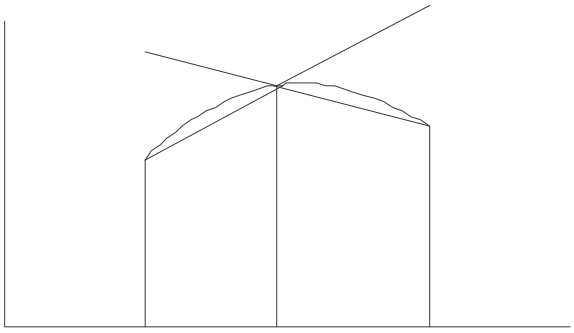


Рис. 8.62. Получение параболы из прямых

а на отрезке $[x_i, x_{i+1}]$ – локальная функция

$$f_i(x) = y_i + \frac{y_{i+1} - y_i}{x_{i+1} - x_i}(x - x_i) \quad .$$

Продолжим функции $f_{i-1}(x)$ и $f_i(x)$ таким образом, чтобы каждая из них была определена на отрезке $[x_{i-1}, x_{i+1}]$ (рис. 8.62).

Используя ту же формулу (8.148) для представления функции на отрезке $[x_{i-1}, x_{i+1}]$, получим выражение

$$F_{i-1}(x) = y_{i-1} + \frac{y_i - y_{i-1}}{x_i - x_{i-1}}(x - x_{i-1}) +$$

$$+ \frac{y_i - y_{i-1} + \frac{y_{i+1} - y_i}{x_{i+1} - x_i}(x - x_i) + \frac{y_i - y_{i-1}}{x_i - x_{i-1}}(x - x_{i-1})}{x_{i+1} - x_{i-1}}(x - x_{i-1}). \quad (8.149)$$

Подстановкой в данное выражение значений x_{i-1}, x_i и x_{i+1} вместо x можно убедиться в том, что выполняются соотношения

$$\left.\begin{aligned} F_{i-1}(x_{i-1}) &= y_{i-1} \\ F_{i-1}(x_i) &= y_i \\ F_{i-1}(x_{i+1}) &= y_{i+1} \end{aligned}\right\}.$$

Следовательно, полученная функция проходит через точки P_{i-1}, P_i и P_{i+1}.

Полученное выражение мы можем записать также как

$$F_i(x) = y_{i-1} + \frac{y_i - y_{i-1}}{x_i - x_{i-1}}(x - x_{i-1}) +$$

$$+ \frac{y_i - y_{i-1} + \frac{y_{i+1} - y_i}{x_{i+1} - x_i}(x - x_i) + \frac{y_i - y_{i-1}}{x_i - x_{i-1}}(x - x_{i-1})}{x_{i+1} - x_{i-1}}(x - x_{i-1}). \quad (8.150)$$

По существу, это вопрос выбора обозначений. Но последнее обозначение представляется более логичным. Оно позволяет ввести единообразие, считая, что каждая функция задана в узлах, а во внутренних точках отрезка $[x_i, x_{i+1}]$ определение интерполирующей функции осуществляется динамически.

Заменив формулу для интерполирования вперед ее эквивалентом – формулой для интерполирования назад

$$y_{i-1} + \frac{y_i - y_{i-1}}{x_i - x_{i-1}}(x - x_{i-1}) = y_i + \frac{y_{i-1} - y_i}{x_{i-1} - x_i}(x - x_i),$$

получаем соотношение

$$F_i(x) = y_{i-1} + \frac{y_i - y_{i-1}}{x_i - x_{i-1}}(x - x_{i-1}) +$$

$$+ \frac{\frac{y_{i+1} - y_i}{x_{i+1} - x_i}(x - x_i) - \frac{y_{i-1} - y_i}{x_{i-1} - x_i}(x - x_i)}{x_{i+1} - x_{i-1}}(x - x_{i-1}). \quad (8.151)$$

Линейное преобразование. Парабола $y = x^2$, заданная своими точками $p_{i-1} = (-1,1)$, $p_i = (0,0)$ и $p_{i+1} = (1,1)$, с помощью линейного преобразования может быть преобразована в кривую, проходящую через точки P_{i-1}, P_i и P_{i+1}. Предварительно выполним преобразование системы координат: начало перенесем в точку P_i, а оси развернем на угол α, так что $Y_{i-1} = Y_{i+1}$. Тогда задача сводится к преобразованию плоскости таким образом, чтобы точка p_j параболы перешла в точку P_j $(j = i-1, i, i+1)$ (рис. 8.63).

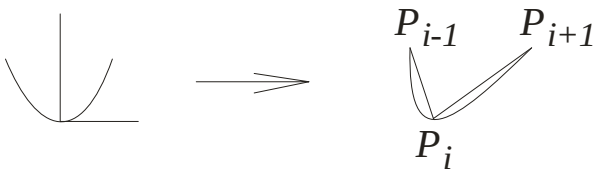


Рис. 8.63. Линейное преобразование

Такое преобразование будем осуществлять по формулам

$$\left.\begin{aligned} a_0 + a_1 x + a_2 y &= X \\ b_0 + b_1 x + b_2 y &= Y \end{aligned}\right\}, \qquad (8.152)$$

где x и y – координаты точки-прообраза, а X и Y – координаты образа точки. Так как $y = x^2$, мы можем записать эти соотношения в виде

$$\left.\begin{aligned} a_0 + a_1 x + a_2 x^2 &= X \\ b_0 + b_1 x + b_2 x^2 &= Y \end{aligned}\right\}. \qquad (8.153)$$

Коэффициенты первого уравнения могут быть найдены из условий

$$a_0 + a_1 x_j + a_2 x_j^2 = X_j \quad (j = i-1, i, i+1).$$

Тогда

$$\left.\begin{aligned} a_0 - a_1 + a_2 &= X_{i-1} \\ a_0 &= 0 \\ a_0 + a_1 + a_2 &= X_{i+1} \end{aligned}\right\},$$

и из решения этой системы уравнений находим

$$\left.\begin{aligned} a_0 &= 0 \\ a_1 &= \frac{X_{i+1} - X_{i-1}}{2} \\ a_2 &= \frac{X_{i-1} + X_{i+1}}{2} \end{aligned}\right\}. \qquad (8.154)$$

Аналогично из условий

$$\left.\begin{aligned} b_0 - b_1 + b_2 &= Y_{i-1} \\ b_0 &= 0 \\ b_0 + b_1 + b_2 &= Y_{i-1} \end{aligned}\right\}$$

получаем коэффициенты

$$\left.\begin{aligned} b_0 &= 0 \\ b_1 &= 0 \\ b_2 &= Y_{i-1} \end{aligned}\right\}. \qquad (8.155)$$

Тогда преобразование параболы может осуществляться по формулам

$$\left.\begin{aligned} a_1 x + a_2 x^2 &= X \\ b_2 x^2 &= Y \end{aligned}\right\},$$

где x играет роль параметра. Чтобы выразить зависимость между X и Y, из второго уравнения находим значение x

$$x = \sqrt{\frac{Y}{b_2}},$$

которое подставляем в первое уравнение и получаем

$$a_1 \sqrt{\frac{Y}{b_2}} = X - a_2 \frac{Y}{b_2}.$$

После возведения в квадрат левой и правой части приходим к выражению

$$a_1^2 \frac{Y}{b_2} = (X - a_2 \frac{Y}{b_2})^2,$$

после преобразования которого получаем уравнение кривой, проходящей через точки P_{i-1}, P_i и P_{i+1}

$$a_2^2 Y^2 - b_2 (2a_2 X + a_1^2) Y + b_2^2 X^2 = 0. \qquad (8.156)$$

Проверить факт прохождения кривой через три заданные точки можно подстановкой их координат в полученное уравнение.

Будем рассматривать полученное выражение как квадратное уравнение относительно Y. Тогда

$$Y = \frac{b_2}{2a_2^2} (2a_2 X + a_1^2 \pm a_1 \sqrt{4a_2 X + a_1^2}). \qquad (8.157)$$

Чтобы определить знак перед радикалом, воспользуемся тем, что при $X = 0$ и $Y = 0$. Следовательно,

$$Y = \frac{b_2}{2a_2^2}(a_1^2 \pm a_1\sqrt{a_1^2}) = 0$$

.

Но это возможно только в том случае, когда перед знаком радикала стоит знак «минус». Поэтому формула (8.157) получает окончательный вид

$$Y = \frac{b_2}{2a_2^2}(2a_2X + a_1^2 - a_1\sqrt{4a_2X + a_1^2}) \quad . \qquad (8.158)$$

Производная полученной функции

$$Y' = \frac{b_2}{a_2}(1 - \frac{a_1}{\sqrt{4a_2X + a_1^2}}) \qquad (8.159)$$

при $X = 0$ также равна 0.

Функция имеет вторую производную

$$Y'' = -\frac{2a_1b_2}{(4a_2X + a_1^2)^{3/2}}, \qquad (8.160)$$

а ее кривизна выражается формулой

$$k = -\frac{2a_1b_2}{(4a_2X + a_1^2)^{3/2}[1 + \frac{b_2^2}{a_2^2}(1 - \frac{a_1}{\sqrt{4a_2X + a_1^2}})^2]^{3/2}} \quad . \qquad (8.161)$$

Коэффициенты полинома второй степени, проходящего через три заданные точки, однозначно определяются из решения системы линейных уравнений с тремя неизвестными. Чтобы иметь возможность управлять поведением кривой, необходим дополнительный параметр. Его можно получить, если повысить степень интерполирующего полинома и использовать функцию

$$y = a_0 + a_1x + a_2x^2 + a_3x^3 \quad . \qquad (8.162)$$

С целью упрощений начало системы координат перенесем в точку P_i. Коэффициенты полинома определим из условий прохождения его через три заданные точки и дополнительного условия, в соответствии с которым, интерполирующая функция в точке P_i имеет максимальное отклонение от хорды $[P_{i-1}, P_{i+1}]$. Данное требование эквивалентно условию параллельности касательной в этой точке и хорды $[P_{i-1}, P_{i+1}]$. Следовательно, должны выполняться соотношения

$$\left.\begin{aligned} &a_0 + a_1 x_{i-1} + a_2 x_{i-1}^2 + a_3 x_{i-1}^3 = y_{i-1} \\ &a_0 = 0 \\ &a_0 + a_1 x_{i+1} + a_2 x_{i+1}^2 + a_3 x_{i+1}^3 = y_{i+1} \\ &a_1 = \frac{y_{i+1} - y_{i-1}}{x_{i+1} - x_{i-1}} \end{aligned}\right\}.$$

Коэффициенты a_2 и a_3 определяются из решения двух линейных уравнений с двумя неизвестными

$$\left.\begin{aligned} &a_1 x_{i-1} + a_2 x_{i-1}^2 + a_3 x_{i-1}^3 = y_{i-1} \\ &a_1 x_{i+1} + a_2 x_{i+1}^2 + a_3 x_{i+1}^3 = y_{i+1} \end{aligned}\right\}.$$

Решив эту систему, получим

$$a_2 = \frac{(y_{i-1} - a_1 x_{i-1}) x_{i+1}^3 - (y_{i+1} - a_1 x_{i+1}) x_{i-1}^3}{(x_{i+1} - x_{i-1}) x_{i-1}^2 x_{i+1}^2},$$

$$a_3 = \frac{(y_{i-1} - a_1 x_{i-1}) x_{i+1}^2 - (y_{i+1} - a_1 x_{i+1}) x_{i-1}^2}{(x_{i-1} - x_{i+1}) x_{i-1}^2 x_{i+1}^2}.$$

Если ввести обозначения

$$\left.\begin{aligned} p &= (y_{i-1} - a_1 x_{i-1}) x_{i+1}^2 \\ q &= (y_{i+1} - a_1 x_{i+1}) x_{i-1}^2 \\ r &= (x_{i-1} - x_{i+1}) x_{i-1}^2 x_{i+1}^2 \end{aligned}\right\},$$

то коэффициенты будут равны

$$\left.\begin{aligned} a_2 &= \frac{p x_{i+1} - q x_{i-1}}{r} \\ a_3 &= \frac{p - q}{r} \end{aligned}\right\}.$$

Параметрическое представление. С течением времени для интерполирования кривых все чаще стало применяться параметрическое представление, когда переменные x и y являются функциями некоторого параметра t:

$$\left.\begin{aligned} x &= x(t) \\ y &= y(t) \end{aligned}\right\}.$$

Используемые при этом функции, как правило, представляют собой полиномы невысокой степени. Примером могут служить полиномы третьей степени:

$$\left.\begin{aligned} x &= a_0 + a_1 t + a_2 t^2 + a_3 t^3 \\ y &= b_0 + b_1 t + b_2 t^2 + b_3 t^3 \end{aligned}\right\}. \qquad (8.163)$$

Использование параметрического представления полностью снимает проблемы, связанные с неоднозначностью функций, применяемых для интерполирования кривых. При выборе параметра наиболее естественным решением является использование в качестве такового длины кривой. Но поскольку кривая неизвестна, а, следовательно, и ее длина, то длину кривой заменяют длиной ломаной, представляющей кривую.

Наряду с таким решением возможно другое, когда в качестве параметра используется номер (индекс) точки. Нумерация точек может рассматриваться как два отображения множества целых чисел на множество вещественных чисел

$$I_x : (1, ..., n) \to (x_1, ..., x_n),$$
$$I_y : (1, ..., n) \to (y_1, ..., y_n),$$

ставящие в соответствие номеру точки ее координаты. Преимуществом такого подхода является то, что формулы имеют более простой вид.

Рассмотрим интерполирование кривой с помощью соотношений (8.163), когда параметр t является номером точки. Нумерация точек на замкнутой кривой содержит элемент случайности. Мы можем начинать нумерацию с любой точки и с любого целого числа. Важно лишь то, что номер каждой следующей точки больше номера предыдущей на 1. Поэтому, чтобы сделать вывод более простым, введем обозначения, представленные на рис. 8.64, то есть, условимся считать, что три последовательно расположенные точки имеют номера –1, 0 и 1.

Тогда коэффициенты уравнения

$$x = a_0 + a_1 t + a_2 t^2 + a_3 t^3 \qquad (8.164)$$

могут быть получены из условий прохождения функции через три исходные точки

$$\left.\begin{aligned} a_0 - a_1 + a_2 - a_3 &= x_{-1} \\ a_0 &= x_0 \\ a_0 + a_1 + a_2 + a_3 &= x_1 \end{aligned}\right\} \qquad (8.165)$$

Данная система уравнений

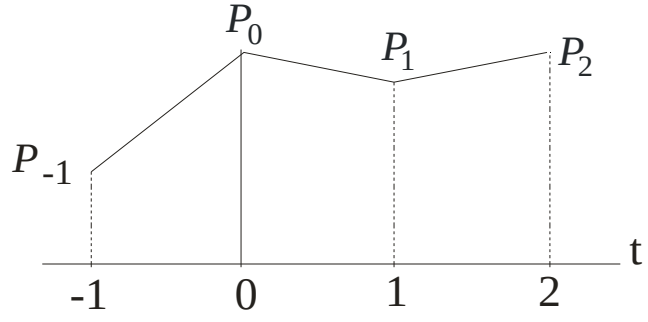


Рис. 8.64. Параметрическое представление

содержит 4 неизвестных. В качестве четвертого условия примем, что касательная к кривой в узле $t = 0$ параллельна хорде $[P_{-1}, P_1]$. Следовательно, производная функции

$$\frac{dx}{dt} = a_1 + 2a_2 t + 3a_3 t^2$$

в точке $t = 0$ равна $\frac{x_1 - x_{-1}}{2}$, и тогда

$$a_1 = \frac{x_1 - x_{-1}}{2}.$$

Если дополним полученную выше систему уравнений (8.165) последним равенством, то мы приведем ее к системе двух линейных уравнений с двумя неизвестными

$$\left.\begin{aligned} a_2 - a_3 &= x_{-1} - x_0 + a_1 \\ a_2 + a_3 &= x_1 - x_0 - a_1 \end{aligned}\right\},$$

из решения которой находим

$$\left.\begin{aligned} a_2 &= \frac{x_{-1} + x_1}{2} - x_0 \\ a_3 &= 0 \end{aligned}\right\}.$$

Тогда уравнения (8.164) примут вид

$$\left.\begin{aligned} x &= a_0 + a_1 t + a_2 t^2 \\ y &= b_0 + b_1 t + b_2 t^2 \end{aligned}\right\}. \qquad (8.165.1)$$

Вообще-то, мы не предполагали равенства $a_3 = 0$, поэтому воспримем его как вознаграждение за усилия. Отсюда следует, что при проведении параболы (8.165.1) через три точки при $t = 0$ обеспечивается параллельность касательной и стягивающей хорды. Таким образом, эти требования являются эквивалентными.

Чтобы проверить этот факт, проведем параболу (8.165.1) через три точки и определим значение ее производной при $t = 0$. Из условий

$$\left.\begin{aligned} a_0 - a_1 + a_2 &= x_{-1} \\ a_0 &= x_0 \\ a_0 + a_1 + a_2 &= x_1 \end{aligned}\right\}$$

находим значения коэффициентов

$$\left.\begin{aligned} a_0 &= x_0 \\ a_1 &= \frac{x_1 - x_{-1}}{2} \\ a_2 &= \frac{x_{-1} + x_1}{2} - x_0 \end{aligned}\right\}.$$

Производная функции

$$\frac{dx}{dt} = a_1 + 2a_2 t$$

при $t = 0$ равна a_1, а это и есть тангенс хорды.

Для коэффициентов b могут быть получены аналогичные выражения

$$\left.\begin{aligned} b_0 &= y_0 \\ b_1 &= \frac{y_1 - y_{-1}}{2} \\ b_2 &= \frac{y_{-1} + y_1}{2} - y_0 \end{aligned}\right\}. \qquad (8.166)$$

Чтобы найти значение производной $\frac{dy}{dx}$ в точке $t = 0$, воспользуемся соотношением

$$\frac{dy}{dx} = \frac{\frac{dy}{dt}}{\frac{dx}{dt}},$$

из которого находим

$$\frac{dy}{dx} = \frac{b_1 + 2b_2 t}{a_1 + 2a_2 t}. \qquad (8.167)$$

Подставляя в данное выражение $t = 0$, получим значение производной в средней точке

$$\left(\frac{dy}{dx}\right)_{t=0} = \frac{b_1}{a_1}$$

или

$$\left(\frac{dy}{dx}\right)_{t=0} = \frac{y_1 - y_{-1}}{x_1 - x_{-1}}.$$

Таким образом, используя параметрическое представление (8.165), мы получили кривую, которая проходит через три заданные точки. Используемые для ее построения полиномы имеют невысокую степень, и их коэффициенты

вычисляются по очень простым формулам. Касательная к кривой в средней точке параллельна хорде между крайними точками. Такая задача первоначально не ставилась, но поскольку такое свойство является желательным, будем относиться к нему как к подарку. Полученный результат можно считать достаточно хорошим.

Проведем теперь кривую через точки 0, 1 и 2. Но при этом из практических соображений будем использовать не формулу (8.165.1), а соотношения

$$\left.\begin{aligned} x &= a_0 + a_1(t-1) + a_2(t-1)^2 \\ y &= b_0 + b_1(t-1) + b_2(t-1)^2 \end{aligned}\right\}.$$

Из условий

$$\left.\begin{aligned} a_0 - a_1 + a_2 &= x_0 \\ a_0 &= x_1 \\ a_0 + a_1 + a_2 &= x_2 \end{aligned}\right\}$$

определяются коэффициенты первого уравнения

$$\left.\begin{aligned} a_0 &= x_1 \\ a_1 &= \frac{x_2 - x_0}{2} \\ a_2 &= \frac{x_0 + x_2}{2} - x_1 \end{aligned}\right\}. \qquad (8.168)$$

Таким же образом находим коэффициенты второго уравнения

$$\left.\begin{aligned} b_0 &= y_1 \\ b_1 &= \frac{y_2 - y_0}{2} \\ b_2 &= \frac{y_0 + y_2}{2} - y_1 \end{aligned}\right\}. \qquad (8.169)$$

В результате подстановки значений коэффициентов и значений параметра t в исходные уравнения можно убедиться, что кривая проходит через точки 1, 2 и 3.

Теперь мы можем получить уравнение кривой на отрезке [0, 1], для чего воспользуемся формулой линейной интерполяции (8.148)

$$F_i(x) = f_i(x) + \frac{f_{i+1}(x) - f_i(x)}{x_{i+1} - x_i}(x - x_i).$$

Тогда, поскольку $t_0 = 0$, нужная формула будет иметь вид

$$x = x^{(0)} + (x^{(1)} - x^{(0)})t$$

и

$$x = a_0^{(0)} + a_1^{(0)}t + a_2^{(0)}t^2 + [a_0^{(1)} + a_1^{(1)}(t-1) + a_2^{(1)}(t-1)^2 - a_0^{(0)} - a_1^{(0)}t - a_2^{(0)}t^2]t$$

.

После подстановки значений коэффициентов уравнение можно записать как

$$x = x_0 + (\frac{x_1 - x_{-1}}{2})t + (\frac{2x_{-1} - 5x_0 + 4x_1 - x_2}{2})t^2 +$$

$$+ (\frac{-x_{-1} + 3x_0 - 3x_1 + x_2}{2})t^3. \qquad (8.170)$$

Аналогичные выражения для y имеют вид

$$y = y_0 + (\frac{y_1 - y_{-1}}{2})t + (\frac{2y_{-1} - 5y_0 + 4y_1 - y_2}{2})t^2 +$$

$$+ (\frac{-y_{-1} + 3y_0 - 3y_1 + y_2}{2})t^3. \qquad (8.171)$$

Данный вывод интерполяционных формул является довольно сложным. Можно получить более простой способ интерполирования кривых, основанный на параметрическом представлении. Пусть кривая описывается уравнениями

$$x = a_0 + a_1t + a_2t^2 + a_3t^3;$$

$$y = b_0 + b_1t + b_2t^2 + b_3t^3.$$

Ее производные по параметру t имеют вид

$$x' = a_1 + 2a_2t + 3a_3t^2;$$

$$y' = b_1 + 2b_2t + 3b_3t^2.$$

Коэффициенты можно определить по значениям функции и ее производных на концах отрезка. Тогда для отрезка [0, 1] (см. рис. 8.64) получим систему уравнений

$$a_0 = x_0;$$

$$a_1 = k_0;$$

$$a_0 + a_1 + a_2 = x_1;$$

$$a_1 + 2a_2 + 3a_3 = k_1,$$

где k_0 и k_1 – значения производной в точках 0 и 1, вычисляемые как

$$k_0 = \frac{x_1 - x_{-1}}{2};$$

$$k_1 = \frac{x_2 - x_0}{2}.$$

Следовательно,

$$a_1 = \frac{x_1 - x_{-1}}{2}.$$

Исключив два известных коэффициента a_0 и a_1, приходим к двум уравнениям с двумя неизвестными

$$\left.\begin{aligned} a_2 + a_3 &= x_1 - x_0 - \frac{x_1 - x_{-1}}{2} \\ 2a_2 + 3a_3 &= \frac{x_2 - x_0}{2} - \frac{x_1 - x_{-1}}{2} \end{aligned}\right\}.$$

Отсюда находим значения коэффициентов

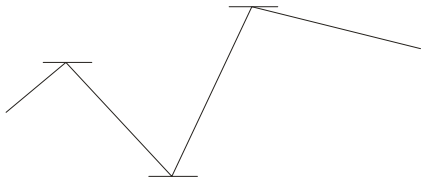
$$\left.\begin{aligned} a_2 &= \frac{2x_{-1} - 5x_0 + 4x_1 - x_2}{2} \\ a_3 &= \frac{x_2 - 3x_1 + 3x_0 - x_{-1}}{2} \end{aligned}\right\}. \qquad (8.172)$$

Таким образом, мы получили те же формулы для интерполирования кривых с помощью параметрического представления, хотя исходные посылки были разные.

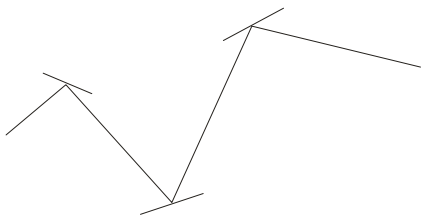
Приведенные выше способы интерполирования функций одной переменной и кривых ни в коей мере не исчерпывают всех возможностей, а служат лишь иллюстрацией различий в исходных положениях, разных подходов к решению этой задачи. В частности, не рассматривались методы получения интерполирующих функций с непрерывной второй производной, поскольку способы их построения достаточно очевидны, а необходимость в них возникает редко. Содержание данной главы можно рассматривать, скорее всего, как введение в проблему интерполирования кривых.

Об интерполировании кривых в целом, видимо, можно сказать, что не существует способа ни наилучшего, ни универсального. Косвенным доказательством этого утверждения служат если не тысячи, то сотни публикаций по методам интерполирования функций одной переменной и кривых. Их авторы находят достаточно оснований для разработки и использования новых методов.

Чтобы подтвердить данное утверждение наглядно, обратимся к рис. 8.65, на котором две ломаных линии представляют собой один и тот же набор точек. Если данные точки принадлежат профилю, то направления касательных должны быть выбраны так, как указано на рис. 8.65, а, поскольку они являются точками локальных экстремумов. Если эти же



а)



б)

Рис. 8.65. Направления касательных

точки являются точками на горизонтали, то направления касательных должны соответствовать указанным на рис. 8.65, б.

Отсюда следует, что значения производной интерполирующей функции в исходных точках должны вычисляться с учетом физических соображений. Можно также сделать вывод, что для достижения универсальности программного обеспечения (если это можно так назвать), необходимо иметь несколько способов определения значений производных, реализованных в виде отдельных модулей или процедур, и вместе с координатами исходных точек хранить значения производных в них. Другими словами, и в данном случае разумно придерживаться принципов функциональной избыточности и функциональной избирательности. И, конечно, необходимы такие средства, как визуальный контроль и корректировка вычисленных значений производных (то есть направлений касательных) в интерактивном режиме. Но универсальным такой способ можно назвать лишь условно. На этой пессимистической ноте мы и закончим рассмотрение интерполяции кривых.

8.18. Методы восполнения регулярных моделей

Поскольку значения высот в дискретных моделях топографических поверхностей заданы только на конечном множестве точек, а при решении задач требуется знать значение высоты в любой точке поверхности, постольку необходимы методы вычисления высоты в произвольной точке по конечному множеству точек на поверхности. Методы создания непрерывных моделей рассматривать не будем как не оправдавшие надежд. Тогда эта задача может быть названа задачей восполнения дискретных моделей до кусочно-непрерывных.

Среди дискретных моделей самыми простыми по своей структуре являются модели на сетке прямоугольников или квадратов (рис. 8.66). Чаще всего задача вычисления в произвольной точке значения функции двух переменных, заданной своими значениями в узлах регулярной сетки, решается с применением билинейных и бикубических сплайнов, являющихся обобщением линейных и кубических сплайнов на случай функций двух переменных.

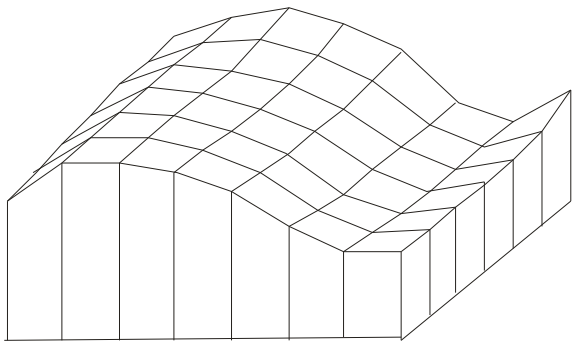


Рис. 8.66. Регулярная модель

Наиболее простым способом восстановления значения функции являются билинейные сплайны. Для этого на каждом прямоугольном элементе вводится локальная система координат таким образом, что выполняются соотношения $x \in [0,1]$ и $y \in [0,1]$ (рис. 8.67).

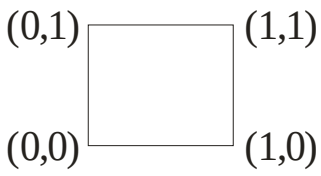


Рис. 8.67. Локальные координаты

Поверхность на каждом таком элементе описывается билинейным уравнением

$$f(x,y) = a_0 + a_1 x + a_2 y + a_3 xy. \quad (8.173)$$

Коэффициенты данного уравнения могут быть получены из условий прохождения поверхности через исходные точки:

$$\left.\begin{aligned} f(0,0) &= a_0, \\ f(1,0) &= a_0 + a_1, \\ f(0,1) &= a_0 + a_2, \\ f(1,1) &= a_0 + a_1 + a_2 + a_3 \end{aligned}\right\}. \quad (8.174)$$

Отсюда следуют формулы для получения коэффициентов билинейного уравнения

$$\left.\begin{aligned} a_0 &= f(0,0), \\ a_1 &= f(1,0) - a_0, \\ a_2 &= f(0,1) - a_0, \\ a_3 &= f(1,1) - a_0 - a_1 - a_2 \end{aligned}\right\}. \quad (8.175)$$

Сечение полученной поверхности вертикальной плоскостью, параллельной оси абсцисс или ординат, является отрезком прямой линии. Например, при фиксированном значении $y = c$ получаем уравнение прямой

$$f(x,c) = (a_0 + a_2 c) + (a_1 + a_3 c)x.$$

Функция, описывающая поверхность в целом, является непрерывной, но не гладкой, ее первые производные терпят разрыв на границах соседних элементов.

При использовании бикубических сплайнов поверхность на каждом прямоугольном элементе представляется уравнением

$$\begin{aligned} h = {} & a_{00} + a_{01}y + a_{02}y^2 + a_{03}y^3 + \\ & + a_{10}x + a_{11}xy + a_{12}xy^2 + a_{13}xy^3 + \\ & + a_{20}x^2 + a_{21}x^2 y + a_{22}x^2 y^2 + a_{23}x^2 y^3 + \\ & + a_{30}x^3 + a_{31}x^3 y + a_{32}x^3 y^2 + a_{33}x^3 y^3. \end{aligned} \quad (8.176)$$

Локальные бикубические сплайны строятся следующим образом. В узлах прямоугольной сетки задаются значения функции $f(x,y)$ и ее производных $f_x^{'}(x,y)$, $f_y^{'}(x,y)$, $f_{xy}^{''}(x,y)$. Тогда для каждого прямоугольного элемента поверхности известны 16 значений, и для определения 16 коэффициентов каждого бикубического уравнения составляется система из 16 линейных уравнений.

Бикубические сплайны на сетке прямоугольников, как и описанные далее способы, находят применение в тех случаях, когда тем или иным образом

получены значения функции в ее узлах, но сетка является недостаточно плотной и требуется осуществить ее сгущение.

Судя по сообщениям о попытках приближения топографических поверхностей бикубическими сплайнами, более точными оказываются результаты, полученные с помощью локальных сплайнов [27]. Интуитивно таких выводов следовало ожидать, если учесть, что топографические поверхности обладают плохими дифференциальными свойствами. Поэтому рассматривать полные бикубические сплайны не будем, так как локальные бикубические сплайны обеспечивают более высокую точность представления топографических поверхностей.

С методом локальных сплайнов связаны так называемые *поверхности Кунса*. Хотя они применялись, вероятно, только в техническом конструировании, их рассмотрение здесь представляется целесообразным по следующим причинам. Во-первых, в работе [26, с. 172] они названы «обобщением обычных методов представления поверхностей», а в [14, с. 5] Ю.С. Завьялов характеризует их как «общий способ построения по заданному каркасу поверхности». Методы, обладающие таким свойством, как универсальность, особенно привлекательны при разработке программного обеспечения. Во-вторых, поверхности Кунса могут использоваться для представления естественных и искусственных (существующих или проектируемых) топографических поверхностей.

Сущность метода Кунса заключается в следующем (рис. 8.68). На поверхности выбираются два семейства кривых u и v так, что кривые одного семейства не пересекаются между собой, а две кривые из разных семейств пересекаются только в одной точке. Выбранные линии образуют каркас и служат в качестве координатных линий $u=const$ и $v=const$ *криволинейных координат* u и v. Такая поверхность гомеоморфна прямоугольной области на плоскости. Другими словами, участок поверхности, ограниченный координатными линиями $u=u_1$, $u=u_m$ и $v=v_1$, $v=v_n$, может быть трансформирован в плоский прямоугольник и обратно.

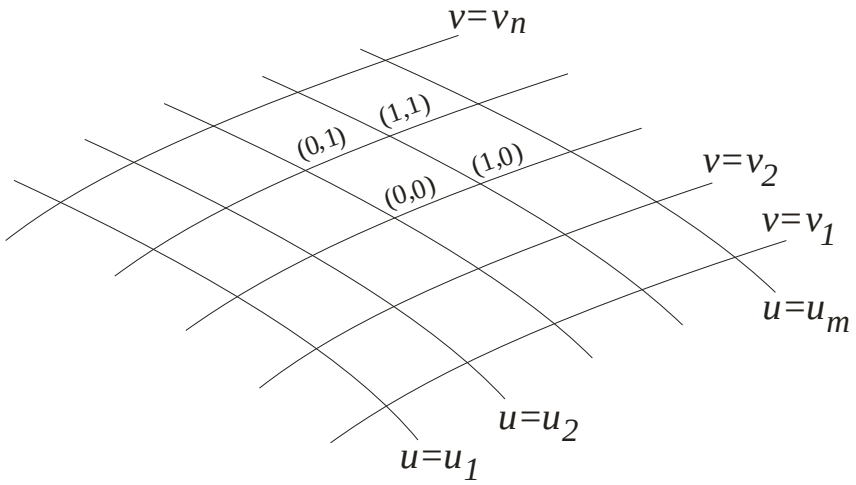


Рис. 8.68. Поверхность Кунса

Поэтому, не теряя общности в рассуждениях, в дальнейшем удобно считать, что линии каркаса образуют прямоугольную сеть на плоскости, и

вдоль каждой стороны прямоугольного элемента поверхность совпадает с кривыми $f_i(u, v_i)$ $(i = 1, ..., n)$ и $f_j(u_j, v)$ $(j = 1, ..., m)$.

Для каждого прямоугольного элемента вводится такая локальная система координат, что в его пределах координаты u и v изменяются от 0 до 1. Тогда поверхность на прямоугольном элементе $u = 0$, $u = 1$, $v = 0$, $v = 1$ определяется выражением (в нотации Принса):

$$H(u,v) = (0,v)F_0(u) + (1,v)F_1(u) + (u,0)F_0(v) + (u,1)F_1(v) -$$

$$-(0,0)F_0(u)F_0(v) - (0,1)F_0(u)F_1(v) - (1,0)F_1(u)F_0(v) - (1,1)F_1(u)F_1(v),$$

где $H(u,v)$ – приближающая функция; $(0,v)$, (1, v), (u, 0), $(u,1)$ – граничные кривые прямоугольника; (0, 0), (0, 1), (1, 0), (1, 1) – значения функции в углах прямоугольника; $F_0(u)$, $F_1(u)$, $F_0(v)$, $F_1(v)$ – так называемые «функции сшивания».

Очевидно, что поверхность Кунса определяется неоднозначно, так как зависит от выбора функций сшивания. В простейшем случае, когда о приближаемой поверхности известно только то, что она непрерывна, достаточно, чтобы функции сшивания отвечали условиям: $F_0(0) = 1$, $F_0(1) = 0$, $F_1(0) = 0$, $F_1(1) = 1$. Если требуется еще и непрерывность первых производных интерполирующей функции, то должны выполняться дополнительные условия: $F'_0(0) = 0$, $F'_0(1) = 0$, $F'_1(0) = 0$, $F'_1(1) = 0$. Можно также вывести условия непрерывности производных более высокого порядка, но для моделирования топографических поверхностей это не требуется.

Связь между поверхностями Кунса и сплайнами обсуждалась в [1] и [14]. Чтобы перейти от поверхности Кунса к сплайнам, требуется для каждого прямоугольного элемента, используя уравнения граничных кривых, вычислить значения функции и ее первых производных. Если, кроме того, в узлах прямоугольной сетки задать некоторым произвольным, но фиксированным, способом значения смешанной производной, то для каждого прямоугольника будем иметь 16 неизвестных величин. Особенность поверхностей Кунса в том, что в узлах сетки предполагается равенство смешанных производных нулю, чего обычная техника сплайнов не требует.

8.19. Условия гладкости кусочно-непрерывных функций двух переменных

Дискретная модель топографической поверхности на сетке квадратов, в узлах которой определены значения высот, может оказаться не слишком гладкой и некоторым пользователям потребуется построение более плотной сетки. Такая модель будет не только более гладкой, но может оказаться и более точной. Решение задачи восполнения регулярной кусочно-непрерывной модели на сетке квадратов или прямоугольников может быть получено с помощью метода, представляющего собой обобщение изложенного выше способа

интерполирования функций одной переменной на случай функций двух переменных.

Пусть на плоскости задана прямоугольная область $R: a \le x \le b, c \le y \le d$ и две одномерные сетки

$$\left.\begin{array}{l} \Delta x: a = x_0 < x_1 < \ldots < x_n = b \\ \Delta y: c = y_0 < y_1 < \ldots < y_m = d \end{array}\right\}, \quad (8.177)$$

образующие двумерную сетку

$$\pi: \{P_{ij}\} (i = 0,1, \ldots, n;\ j = 0,1, \ldots, m),$$

которая разбивает прямоугольник R на прямоугольные конечные элементы

$$\{R_{ij}: x_i \le x \le x_{i+1};\ y_j \le y \le y_{j+1}\}\ (i = 0,1,\ldots,n-1;\ j = 0,1,\ldots,m-1).$$

(8.178)

Пусть также приближаемая функция $z(x, y)$ – однозначная, непрерывная и гладкая, то есть, имеющая непрерывные частные производные, – задана своими значениями $z(x_i, y_j)\ (i = 0,1, \ldots, n;\ j = 0,1, \ldots, m)$. Рассмотрим случай, когда приближающая функция $F(x, y)$ должна проходить точно через заданные на приближаемой поверхности точки:

$$F(x_i, y_j) = z(x_i, y_j). \quad (8.179)$$

Интерполирующую функцию будем отыскивать как совокупность кусочно-непрерывных функций

$$F(x, y) = \{F_{ij}(x, y)\}\ (i = 0,1, \ldots, n-1;\ j = 0,1, \ldots, m-1), \quad (8.180)$$

причем областью определения каждой функции $F_{ij}(x, y)$ будет являться соответствующий элемент R_{ij}.

Пусть также в окрестности каждой точки (x_i, y_j) моделируемая функция $z(x, y)$ с достаточной степенью точности может быть представлена локальной функцией $f_{ij}(x, y)$, удовлетворяющей, в частности, соотношению

$$f_{ij}(x_i, y_j) = z(x_i, y_j). \quad (8.181)$$

Тогда по аналогии с интерполированием функций одной переменной функцию $F_{ij}(x, y)$ определим как

$$F_{ij}(x, y) = p_{ij}(x, y) f_{ij}(x, y) + p_{ij+1}(x, y)$$

$$+ p_{i+1\,j}(x, y) f_{i+1\,j}(x, y) + p_{i+1\,j+1}(x, y) \quad (8.182)$$

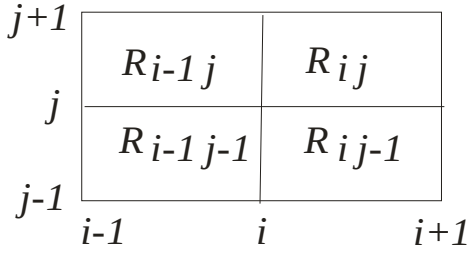


Рис. 8.69. Область определения локальных и весовых функций

Каждая локальная функция $f_{ij}(x,y)$ определена на четырех смежных конечных элементах R_{i-1j-1}, R_{i-1j}, R_{ij-1} и R_{ij}, а областью определения весовых функций $p_{ij}(x,y)$, $p_{ij+1}(x,y)$, $p_{i+1j}(x,y)$ и $p_{i+1j+1}(x,y)$ является конечный элемент R_{ij} (рис. 8.69).

Весовые функции выбираются таким образом, чтобы их сумма отвечала условию

$$p_{ij}(x,y)+p_{ij+1}(x,y)+p_{i+1j}(x,y)+p_{i+1j+1}(x,y)=1, \quad (8.183)$$

то есть, являются «приведенными к единице» весами. Кроме того, каждая весовая функция $p_{ij}(x,y)$ вычисляется как

$$p_{ij}(x,y)=p_i(x)p_j(y), \quad (8.184)$$

где, в свою очередь, соблюдаются соотношения

$$\left.\begin{aligned} p_i(x)+p_{i+1}(x)=1 \\ p_j(y)+p_{j+1}(y)=1 \end{aligned}\right\}. \quad (8.185)$$

С учетом выражения (8.184), интерполирующую функцию $F_{ij}(x,y)$ на элементе R_{ij} можно представить следующим образом:

$$F_{ij}(x,y)=p_i(x)p_j(y)f_{ij}(x,y)+p_i(x)p_{j+1}(y)f_{ij+1}(x,y)+$$

$$+p_{i+1}(x)p_j(y)f_{i+1j}(x,y)+p_{i+1}(x)p_{j+1}(y)f_{i+1j+1}(x,y). \quad (8.186)$$

Очевидно, что для уточненных таким образом весовых функций по-прежнему справедливо условие

$$p_i(x)p_j(y)+p_i(x)p_{j+1}(y)+p_{i+1}(x)p_j(y)+p_{i+1}(x)p_{j+1}(y)=1. \quad (8.187)$$

Если функции $p(x,y)$ и $f(x,y)$ непрерывны на элементе R_{ij}, то и функция (8.186) во всех внутренних точках этого элемента также будет непрерывна. Поведение интерполирующей функции $F(x,y)$ на границе двух соседних элементов не столь очевидно и требует изучения.

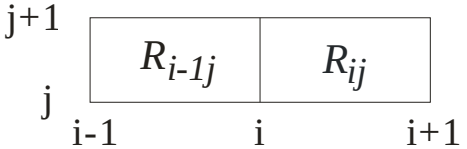


Рис. 8.70. К определению непрерывности

В качестве примера рассмотрим два соседних элемента R_{i-1j} и R_{ij} (рис. 8.70). Непрерывность функции $F(x,y)$ в точке с координатами (x_i,y), где $y\in[y_j,y_{j+1}]$, означает, что

$$F(x_i, y) = F_{i-1\,j}(x_i, y) = F_{ij}(x_i, y) \quad (8.188)$$

Однако, в отличие от интерполирования функций одной переменной (когда значения интерполируемой функции на концах элементов известны), значения функции $f(x,y)$ определены лишь в узлах и не известны на сторонах сетки. Поэтому мы не можем написать выражение, аналогичное выражению для функций одной переменной. Но можно в качестве значения функции $f(x,y)$ взять значение, полученное интерполированием функции одной переменной по узлам, расположенным на прямой $x = x_i$, то есть принять

$$p_{i-1}(x_i)\,p_j(y)\,f_{i-1\,j}(x_i, y) + p_{i-1}(x_i)\,p_{j+1}(y)\,f_{i-1\,j+1}(x_i, y) +$$

$$+ p_i(x_i)\,p_j(y)\,f_{ij}(x_i, y) + p_i(x_i)\,p_{j+1}(y)\,f_{ij+1}(x_i, y) =$$

$$= p_j(y)\,f_{ij}(x_i, y) + p_{j+1}(y)\,f_{ij+1}(x_i, y). \quad (8.189)$$

Заменив в левой части $p_i(x)$ на

$$p_i(x) = 1 - p_{i-1}(x), \quad (8.190)$$

получим

$$p_{i-1}(x_i)\,p_j(y)[\,f_{i-1\,j}(x_i, y) - f_{ij}(x_i, y)] +$$

$$+ p_{i-1}(x_i)\,p_{j+1}(y)[\,f_{i-1\,j+1}(x_i, y) - f_{ij+1}(x_i, y)] = 0. \quad (8.191)$$

Из полученного соотношения видно, что непрерывность интерполирующей функции $F(x,y)$ на границе соседних элементов $R_{i-1\,j}$ и R_{ij} может быть достигнута двумя способами:

– либо выбором таких весовых функций, которые отвечают условиям

$$\left.\begin{aligned} p_{i-1}(x_i) &= 0 \\ p_{i+1}(x_i) &= 0 \end{aligned}\right\}; \quad (8.192)$$

– либо использованием локальных функций, для которых справедливы соотношения

$$\left.\begin{aligned} f_{i-1\,j}(x_i, y) &= f_{ij}(x_i, y) \\ f_{i-1\,j+1}(x_i, y) &= f_{ij+1}(x_i, y) \\ f_{i+1\,j}(x_i, y) &= f_{ij}(x_i, y) \\ f_{i-1\,j+1}(x_i, y) &= f_{ij+1}(x_i, y) \end{aligned}\right\}. \quad (8.193)$$

Если повторить аналогичные рассуждения, например, для элементов R_{ij-1} и R_{ij} и ввести несколько иные обозначения, то условия непрерывности

функции $F(x,y)$ для всех внутренних точек прямоугольника *R* можно записать следующим образом:

1. либо

$$\left.\begin{aligned} p_k(x_i) &= \begin{cases} 1 & (k=i) \\ 0 & (k\neq i) \end{cases} \\ p_l(y_j) &= \begin{cases} 1 & (l=j) \\ 0 & (l\neq j) \end{cases} \end{aligned}\right\}; \qquad (8.194)$$

2. либо

$$\left.\begin{aligned} f_{ij}(x_{i-1},y) &= f_{i-1\,j}(x_{i-1},y) \\ f_{ij}(x_{i+1},y) &= f_{i+1\,j}(x_{i+1},y) \\ f_{ij}(x,y_{j-1}) &= f_{ij-1}(x,y_{j+1}) \\ f_{ij}(x,y_{j+1}) &= f_{ij+1}(x,y_{j+1}) \end{aligned}\right\}, \qquad (8.195)$$

где для первых двух равенств $y\in[y_j,y_{j+1}]$, а для последних двух – $x\in[x_i,x_{i+1}]$.

Часто при восполнении функций требуется не только непрерывность самой функции, но и непрерывность ее первых производных. Непрерывность производной $F'_x(x,y)$ по линии стыковки тех же элементов $R_{i-1\,j}$ и R_{ij} означает

$$\frac{\partial F(x_i,y)}{\partial x}=\frac{\partial F_{i-1\,j}(x_i,y)}{\partial x}=\frac{\partial F_{ij}(x_i,y)}{\partial x}, \qquad (8.196)$$

которое обозначает то же, что и

$$\left.\frac{\partial F(x,y)}{\partial x}\right|_{x=x_i}=\left.\frac{\partial F_{i-1\,j}(x,y)}{\partial x}\right|_{x=x_i}=\left.\frac{\partial F_{ij}(x,y)}{\partial x}\right|_{x=x_i},$$

где $y\in[y_j,y_{j+1}]$.

Продифференцировав (8.186), получим для общей стороны элементов $R_{i-1\,j}$ и R_{ij} следующее выражение:

$$\frac{dp_{i-1}(x_i)}{dx} p_j(y) f_{i-1\,j}(x_i, y) + p_{i-1}(x_i) p_j(y) \frac{\partial f_{i-1\,j}(x_i, y)}{\partial x} +$$

$$+ \frac{dp_{i-1}(x_i)}{dx} p_{j+1}(y) f_{i-1\,j+1}(x_i, y) +$$

$$+ p_{i-1}(x_i) p_{j+1}(y) \frac{\partial f_{i-1\,j+1}(x_i, y)}{\partial x} +$$

$$+ \frac{dp_i(x_i)}{dx} p_j(y) f_{ij}(x_i, y) + p_i(x_i) p_j(y) \frac{\partial f_{ij}(x_i, y)}{\partial x} +$$

$$+ \frac{dp_i(x_i)}{dx} p_{j+1}(y) f_{ij+1}(x_i, y) + p_i(x_i) p_{j+1}(y) \frac{\partial f_{ij+1}(x_i, y)}{\partial x} =$$

$$= p_j(y) \frac{\partial f_{ij}(x_i, y)}{\partial x} + p_{j+1}(y) \frac{\partial f_{ij+1}(x_i, y)}{\partial x}. \qquad (8.197)$$

Здесь и далее выражения вида $\frac{\partial p_i(x_i, y)}{\partial x}$ эквивалентны выражениям $\left.\frac{\partial p_i(x, y)}{\partial x}\right|_{x=x_i}$. Данная условность принята с целью сделать выражения более короткими и обозримыми. Если в соответствии с (8.185) в левой части заменить $p_i(x)$ на $1 - p_{i-1}(x)$, то

$$\frac{dp_{i-1}(x_i)}{dx} p_j(y) f_{i-1\,j}(x_i, y) + p_{i-1}(x_i) p_j(y) \frac{\partial f_{i-1\,j}(x_i, y)}{\partial x} +$$

$$+ \frac{dp_{i-1}(x_i)}{dx} p_{j+1}(y) f_{i-1\,j+1}(x_i, y) + p_{i-1}(x_i) p_{j+1}(y) \frac{\partial f_{i-1\,j+1}(x_i, y)}{\partial x} +$$

$$+ \frac{dp_i(x_i)}{dx} p_j(y) f_{ij}(x_i, y) - p_{i-1}(x_i) p_j(y) \frac{\partial f_{ij}(x_i, y)}{\partial x} +$$

$$+ \frac{dp_i(x_i)}{dx} p_{j+1}(y) f_{ij+1}(x_i, y) - p_{i-1}(x_i) p_{j+1}(y) \frac{\partial f_{ij+1}(x_i, y)}{\partial x} = 0, \qquad (8.198)$$

или иначе

$$p_j(y)[\frac{dp_{i-1}(x_i)}{dx} f_{i-1\,j}(x_i,y) + p_{i-1}(x_i)\frac{\partial f_{i-1\,j}(x_i,y)}{\partial x} +$$

$$+ \frac{dp_i(x_i)}{dx} f_{ij}(x_i,y) - p_{i-1}(x_i)\frac{\partial f_{ij}(x_i,y)}{\partial x}] +$$

$$+ p_{j+1}(y)[\frac{dp_{i-1}(x_i)}{dx} f_{i-1\,j+1}(x_i,y) + p_{i-1}(x_i)\frac{\partial f_{i-1\,j+1}(x_i,y)}{\partial x} +$$

$$+ \frac{dp_i(x_i)}{dx} f_{ij+1}(x_i,y) - p_{i-1}(x_i)\frac{\partial f_{ij+1}(x_i,y)}{\partial x}] = 0. \qquad (8.199)$$

Здесь следует заметить, что $p_j(y) + p_{j+1}(y) = 1$ и, в общем случае, $p_j(y)$ и $p_{j+1}(y)$ не равны нулю. Следовательно, для того чтобы выражение (8.199) было справедливо, необходимо соблюдение двух равенств:

$$\frac{dp_{i-1}(x_i)}{dx} f_{i-1\,j}(x_i,y) + p_{i-1}(x_i)\frac{\partial f_{i-1\,j}(x_i,y)}{\partial x} +$$

$$+ \frac{dp_i(x_i)}{dx} f_{ij}(x_i,y) - p_{i-1}(x_i)\frac{\partial f_{ij}(x_i,y)}{\partial x} = 0;$$

$$\frac{dp_{i-1}(x_i)}{dx} f_{i-1\,j+1}(x_i,y) + p_{i-1}(x_i)\frac{\partial f_{i-1\,j+1}(x_i,y)}{\partial x} +$$

$$+ \frac{dp_i(x_i)}{dx} f_{ij+1}(x_i,y) - p_{i-1}(x_i)\frac{\partial f_{ij+1}(x_i,y)}{\partial x} = 0. \qquad (8.200)$$

Чтобы выяснить условия, при которых выполняются полученные соотношения, необходимо рассмотреть два случая, зависящие от того, каким образом достигалась непрерывность самой функции $F(x,y)$.

1. Пусть непрерывность функции достигалась при помощи ограничений (8.194). Тогда $p_{i-1}(x_i) = 0$, и соотношения (8.200) принимают вид:

$$\left.\begin{aligned} \frac{dp_{i-1}(x_i)}{dx} f_{i-1\,j}(x_i,y) + \frac{dp_i(x_i)}{dx} f_{ij}(x_i,y) = 0 \\ \frac{dp_{i-1}(x_i)}{dx} f_{i-1\,j+1}(x_i,y) + \frac{dp_i(x_i)}{dx} f_{ij+1}(x_i,y) = 0 \end{aligned}\right\}. \qquad (8.201)$$

Из (8.185) следует

$$\frac{dp_i(x)}{dx} = -\frac{dp_{i-1}(x)}{dx}. \qquad (8.202)$$

Тогда (8.201) можно записать как

$$\left.\begin{aligned}\frac{dp_{i-1}(x_i)}{dx}[\,f_{i-1\,j}(x_i,y)-f_{ij}(x_i,y)]=0\\ \frac{dp_{i-1}(x_i)}{dx}[\,f_{i-1\,j+1}(x_i,y)-f_{ij+1}(x_i,y)]=0\end{aligned}\right\}. \qquad (8.203)$$

Равенства (8.203) будут выполняться, если соблюдается одно из условий:

а) локальные функции отвечают требованиям (8.195);

б) на весовые функции наложены дополнительные ограничения

$$\left.\begin{aligned}\frac{dp_{i-1}(x_i)}{dx}=0\\ \frac{dp_i(x_i)}{dx}=0\end{aligned}\right\}. \qquad (8.204)$$

2. Если непрерывность функции $F(x,y)$ была достигнута с помощью условий (8.195), то с учетом (8.202) выражения (8.200) принимают вид

$$\left.\begin{aligned}p_{i-1}(x_i)[\frac{\partial f_{i-1\,j}(x_i,y)}{\partial x}-\frac{\partial f_{ij}(x_i,y)}{\partial x}]=0\\ p_{i-1}(x_i)[\frac{\partial f_{i-1\,j+1}(x_i,y)}{\partial x}-\frac{\partial f_{ij}(x_i,y)}{\partial x}]=0\end{aligned}\right\}. \qquad (8.205)$$

Полученные соотношения будут выполняться, если:

а) весовые функции обладают свойством (8.194);

б) либо локальные функции выбраны так, что

$$\left.\begin{aligned}\frac{\partial f_{i-1\,j}(x_i,y)}{\partial x}=\frac{\partial f_{ij}(x_i,y)}{\partial x}\\ \frac{\partial f_{i-1\,j+1}(x_i,y)}{\partial x}=\frac{\partial f_{ij+1}(x_i,y)}{\partial x}\end{aligned}\right\}. \qquad (8.206)$$

Аналогичным образом могут быть получены условия непрерывности $F_y(x,y)$, $F_{xx}(x,y)$, $F_{yy}(x,y)$ и т. д. Условия непрерывности интерполирующей функции и ее производных представлены в табл. 8.6.

Таблица 8.6. Условия непрерывности $F(x,y)$ и ее производных

Свойства весовых функций	Свойства локальных функций		
	$f_{ij}(x_i, y_j) = h(x_i, y_j)$ (B1)	$\frac{\partial f_{ij}}{\partial x} = \frac{\partial f_{i+1j}}{\partial x}$ $\frac{\partial f_{ij+1}}{\partial x} = \frac{\partial f_{i+1j+1}}{\partial x}$ (B2)	$\frac{\partial^2 f_{ij}}{\partial x^2} = \frac{\partial^2 f_{i+1j}}{\partial x^2}$ $\frac{\partial^2 f_{i+1j}}{\partial x^2} = \frac{\partial^2 f_{i+1j+1}}{\partial x^2}$ (B3)
$p_{ij}(x,y) + p_{i+1j}(x,y) = 1$ $p_{ij+1}(x,y) + p_{i+1j+1}(x,y) = 1$ (A1)		$F(x,y)$	$F'(x,y)$
$\frac{\partial p_{ij}}{\partial x} = \frac{\partial p_{i+1j}}{\partial x}$ $\frac{\partial p_{ij+1}}{\partial x} = \frac{\partial p_{i+1j+1}}{\partial x}$ (A2)	$F(x,y)$	$F'(x,y)$	$F''(x,y)$
$\frac{\partial^2 p_{ij}}{\partial x^2} = \frac{\partial^2 p_{i+1j}}{\partial x^2}$ $\frac{\partial^2 p_{i+1j}}{\partial x^2} = \frac{\partial^2 p_{i+1j+1}}{\partial x^2}$ (A3)	$F'(x,y)$	$F''(x,y)$	$F'''(x,y)$

8.20. Методы восполнения нерегулярных моделей

Задача восполнения нерегулярной модели топографической поверхности возникает в связи с необходимостью определения высоты в любой точке поверхности. В последнее время все большее распространение для моделирования топографических поверхностей (как и в других случаях, связанных с представлением кривых или поверхностей) получает метод конечных элементов (МКЭ). В отечественной практике самой первой такой попыткой, видимо, была работа [11]. Первоначально теория сплайнов и теория метода конечных элементов развивались параллельно и независимо друг от друга, со временем принципы сплайн-функций стали использоваться в методе конечных элементов.

Сущность способа состоит в следующем. Область определения функции разбивается на подобласти, называемые конечными элементами, таким образом, что конечные элементы покрывают всю область, не пересекаются и не вырождаются (линии в точки, четырехугольники в треугольники и т. п.). Интерполирующая функция представляется выражением

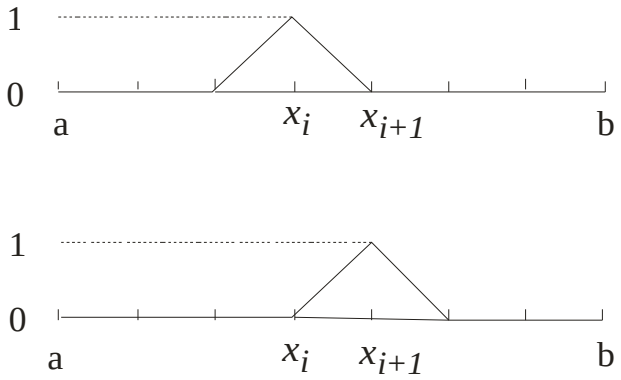


Рис. 8.71. Базисные функции

$$H(x,y)=\sum_{i=1}^{n} f_i(x,y)z_i \quad . \quad (8.207)$$

Функции $f_i(x,y)$ называются базисными, или функциями формы (рис. 8.71, 8.72). Отличие от представления моделируемой функции в виде «обычной» линейной комбинации функций в том, что базисные функции являются финитными. Финитными называют функции, определенные на всей числовой оси, но отличные от нуля лишь на фиксированных конечных элементах. В МКЭ в качестве базисных функций используют сплайны, чаще всего полиномиальные: линейные, билинейные, кубические и т. д. При выборе базисных функций руководствуются формой конечных элементов и необходимой степенью гладкости функции. Свойство финитности обеспечивает разреженную матрицу системы линейных уравнений и устойчивость процесса ее решения.

При едином методологическом подходе метод конечных элементов допускает большое разнообразие вычислительных схем, зависящих от способа разбиения области моделирования и выбора базисных функций. Наиболее удобным с вычислительной точки зрения является случай, когда исходные точки расположены либо в узлах прямоугольной сетки (тогда область стандартно разбивается на прямоугольники или прямоугольные треугольники), либо в узлах сетки равносторонних или равнобедренных треугольников.

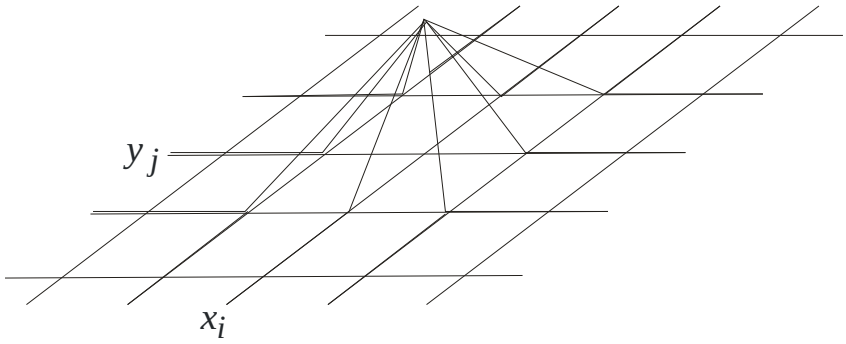


Рис. 8.72. Двумерная базисная функция

Матрицы систем линейных уравнений при этом имеют ленточную структуру, и существуют эффективные алгоритмы решения подобных систем уравнений.

При нерегулярном распределении точек на плоскости матрица хотя и существенно разрежена, но не является, как правило, ленточной; перенумерацией узлов иногда удается привести ее к ленточной структуре.

Базисные функции определяются [15] для треугольных элементов в виде

$$a_{\alpha\beta}x^{\alpha}y^{\beta} \quad (\alpha+\beta=0,...,m), \quad (8.208)$$

а для прямоугольников – как многочлены вида

$$a_{\alpha\beta}x^{\alpha}y^{\beta}\ (\alpha,\beta=0,...,m). \qquad (8.209)$$

В целом метод конечных элементов отличается хорошими интерполяционными свойствами и алгоритмичностью. Рассмотрение МКЭ выходит далеко за пределы данной работы, поскольку даже введение в него может явиться предметом отдельной книги. Дальнейшим обобщением МКЭ является метод граничных элементов, но он находится в стадии становления и для моделирования топографических поверхностей, вероятно, не использовался. Ниже приводятся только самые необходимые начальные сведения из метода конечных элементов.

В двумерном случае наиболее простыми элементами являются треугольники с прямолинейными сторонами, когда интерполяционный полином на треугольнике (рис. 8.73) имеет вид

$$F(x,y)=\alpha_0+\alpha_1x+\alpha_2y. \qquad (8.210)$$

На рис. 8.73 следует обратить внимание на то, что нумерация вершин и их обход совершаются против часовой стрелки.

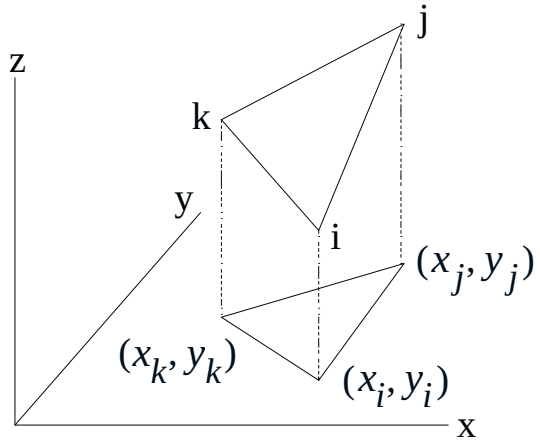


Рис. 8.73. Треугольный элемент

Коэффициенты полинома находятся из условий прохождения плоскости через три вершины треугольника, в результате чего получаем систему линейных уравнений

$$\left.\begin{aligned}\alpha_0+\alpha_1x_i+\alpha_2y_i&=z_i\\ \alpha_0+\alpha_1x_j+\alpha_2y_j&=z_j\\ \alpha_0+\alpha_1x_k+\alpha_2y_k&=z_k\end{aligned}\right\}. \qquad (8.211)$$

Из решения системы находим коэффициенты интерполяционного полинома

$$\left.\begin{aligned}\alpha_0&=\frac{(x_jy_k-x_ky_j)z_i+(x_ky_i-x_iy_k)z_j+(x_iy_j-x_jy_i)z_k}{2S}\\ \alpha_1&=\frac{(y_j-y_k)z_i+(y_k-y_i)z_j+(y_i-y_j)z_k}{2S}\\ \alpha_2&=\frac{(x_k-x_j)z_i+(x_i-x_k)z_j+(x_j-x_i)z_k}{2S}\end{aligned}\right\}, \qquad (8.212)$$

где

$$2S=(y_j-y_k)x_i+(y_k-y_i)x_j+(y_i-y_j)x_k \qquad (8.213)$$

представляет собой площадь проекции треугольника (i, j, k) на плоскость oxy. Эта площадь может также вычисляться по формуле

$$S = \frac{1}{2}\begin{vmatrix} 1 & x_i & y_i \\ 1 & x_j & y_j \\ 1 & x_k & y_k \end{vmatrix}. \qquad (8.214)$$

Подставив полученные значения коэффициентов в (8.210), приходим к выражению

$$F(x,y) = \frac{(x_j y_k - x_k y_j)z_i + (x_k y_i - x_i y_k)z_j + (x_i y_j - x_j y_i)z_k}{2S} +$$
$$+ \frac{(y_j - y_k)z_i + (y_k - y_i)z_j + (y_i - y_j)z_k}{2S} x +$$
$$+ \frac{(x_k - x_j)z_i + (x_i - x_k)z_j + (x_j - x_i)z_k}{2S} y. \qquad (8.215)$$

Перегруппировав члены, данное выражение можно записать как

$$F(x,y) = \frac{(x_j y_k - x_k y_j) + (y_j - y_k)x + (x_k - x_j)y}{2S} z_i +$$
$$+ \frac{(x_k y_i - x_i y_k) + (y_k - y_i)x + (x_i - x_k)y}{2S} z_j +$$
$$+ \frac{(x_i y_j - x_j y_i) + (y_i - y_j)x + (x_j - x_i)y}{2S} z_k. \qquad (8.216)$$

Можно ввести обозначения

$$\left.\begin{aligned} f_i(x,y) &= \frac{(x_j y_k - x_k y_j) + (y_j - y_k)x + (x_k - x_j)y}{2S} \\ f_j(x,y) &= \frac{(x_k y_i - x_i y_k) + (y_k - y_i)x + (x_i - x_k)y}{2S} \\ f_k(x,y) &= \frac{(x_i y_j - x_j y_i) + (y_i - y_j)x + (x_j - x_i)y}{2S} \end{aligned}\right\}, \qquad (8.217)$$

где $f(x,y)$ – функции формы. Тогда выражение (8.216) можно записать коротко в виде

$$F(x,y) = f_i(x,y)z_i + f_j(x,y)z_j + f_k(x,y)z_k. \qquad (8.218)$$

Если ввести обозначения

$$\left.\begin{aligned} a_i &= x_j y_k - x_k y_j \\ b_i &= y_j - y_k \\ c_i &= x_k - x_j \end{aligned}\right\};$$

$$\left.\begin{aligned} a_j &= x_k y_i - x_i y_k \\ b_j &= y_k - y_i \\ c_j &= x_i - x_k \end{aligned}\right\};$$

$$\left.\begin{aligned} a_k &= x_i y_j - x_j y_i \\ b_k &= y_i - y_j \\ c_k &= x_j - x_i \end{aligned}\right\},$$

то функции формы можно представить как

$$\left.\begin{aligned} f_i &= \frac{a_i + b_i x + c_i y}{2S} \\ f_j &= \frac{a_j + b_j x + c_j y}{2S} \\ f_k &= \frac{a_k + b_k x + c_k y}{2S} \end{aligned}\right\}.$$

В вершине i функция формы $f_i(x, y)$ принимает значение

$$f_i(x, y) = \frac{(y_j - y_k)x_i + (y_k - y_i)x_j + (y_i - y_j)x_k}{2S}.$$

Выражение в числителе представляет собой удвоенную площадь треугольника, следовательно,

$$f_i(x_i, y_i) = 1.$$

В вершинах j и k функция формы $f_i(x, y)$ принимает значения

$$f_i(x_j, y_j) = 0;$$

$$f_i(x_k, y_k) = 0.$$

Аналогичным образом можно получить другие равенства, показанные на рис. 8.74.

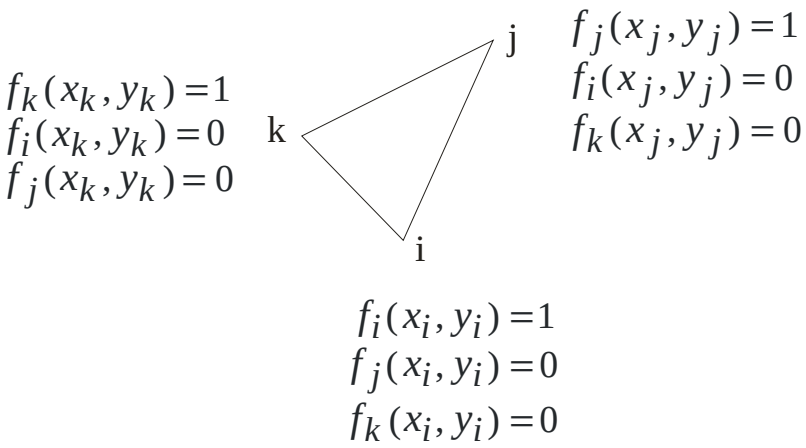


Рис. 8.74. Значения функций формы в вершинах

Кроме того, функция формы каждой вершины принимает нулевые значения в любой точке противолежащей стороны треугольника. Вообще же, значения функции формы изменяются линейно по любому направлению, в том числе между любыми двумя вершинами.

В некоторых случаях более удобной является система локальных координат, начало которой находится в центре треугольника

$$\left.\begin{aligned} x_0 &= \frac{x_i + x_j + x_k}{3} \\ y_0 &= \frac{y_i + y_j + y_k}{3} \end{aligned}\right\}. \qquad (8.219)$$

Локальные координаты обозначим как s и t:

$$\left.\begin{aligned} s &= x - x_0 \\ t &= y - y_0 \end{aligned}\right\}.$$

Заменив в функциях формы координаты x и y на s и t, приходим к выражениям

$$\left.\begin{aligned} f_i(x,y) &= \frac{a_i + b_i x_0 + c_i y_0 + b_i s + c_i t}{2S} \\ f_j(x,y) &= \frac{a_j + b_j x_0 + c_j y_0 + b_j s + c_j t}{2S} \\ f_k(x,y) &= \frac{a_k + b_k x_0 + c_k y_0 + b_k s + c_k t}{2S} \end{aligned}\right\}, \qquad (8.220)$$

из которых видно, что коэффициенты b и c не изменяются. С учетом обозначений для a, b и c, а также выражений (8.213) и (8.219) можно обнаружить, что

$$\left.\begin{aligned} a_i + b_i x_0 + c_i y_0 &= \frac{x_i(y_j - y_k) + x_j(y_k - y_i) + x_k(y_i - y_j)}{3} = \frac{2S}{3} \\ a_j + b_j x_0 + c_j y_0 &= \frac{x_i(y_j - y_k) + x_j(y_k - y_i) + x_k(y_i - y_j)}{3} = \frac{2S}{3} \\ a_k + b_k x_0 + c_k y_0 &= \frac{x_i(y_j - y_k) + x_j(y_k - y_i) + x_k(y_i - y_j)}{3} = \frac{2S}{3} \end{aligned}\right\} .$$

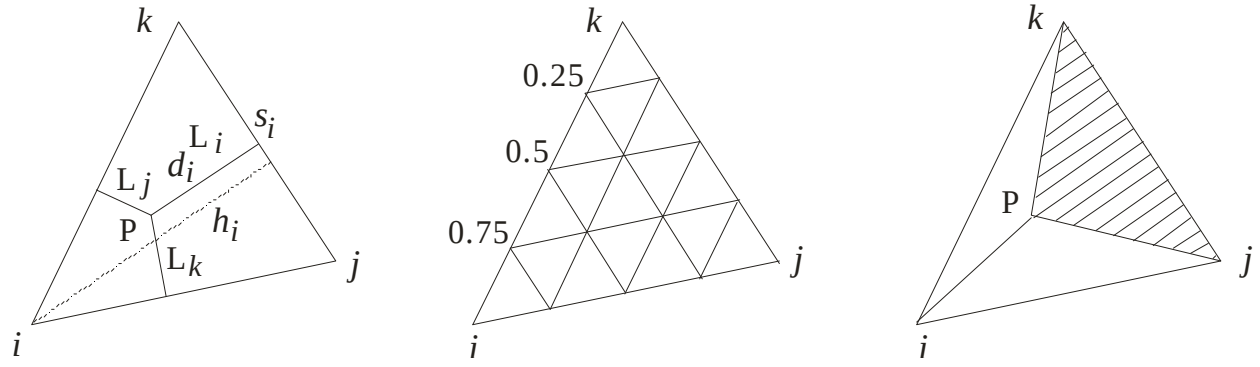
(8.221)

Тогда выражения (8.220) приобретают более простой вид

$$\left.\begin{aligned} f_i(x, y) &= \frac{\frac{2S}{3} + b_i s + c_i t}{2S} \\ f_j(x, y) &= \frac{\frac{2S}{3} + b_j s + c_j t}{2S} \\ f_k(x, y) &= \frac{\frac{2S}{3} + b_k s + c_k t}{2S} \end{aligned}\right\} . \qquad (8.222)$$

Но наиболее часто в приложениях используется система относительных естественных координат, обозначаемых L_i, L_j и L_k и называемых также L-координатами. Координата L_i точки P (рис. 8.75, а) представляет собой отношение расстояния между данной точкой и стороной (j, k) к расстоянию между вершиной i и стороной (j, k), то есть к высоте h_i. Из данного определения следует, что

- координата L_i принимает значения в диапазоне от 0 до 1 $(0 \le L_i \le 1)$;
- линии $L_i = const$ параллельны стороне (j, k) (рис. 8.75, *б*).



а) *б*) *в*)

Рис. 8.75. Система естественных координат треугольника

Подобным же образом определяются координаты L_j и L_k (рис. 8.75, а). Из рис. 8.75.а можно заметить, что L-координаты точки P пропорциональны площади треугольника, образованного этой точкой и двумя другими вершинами исходного треугольника (i, j, k). Рассмотрим, например, значение координаты L_i. Площадь треугольника (i, j, k) может быть вычислена по формуле

$$S_{ijk} = \frac{1}{2} s_i h_i, \qquad (8.223)$$

где h_i – длина перпендикуляра из вершины i на сторону (j, k), а s_i – длина стороны (j, k). Площадь треугольника (P, j, k) равна

$$S_{Pjk} = \frac{1}{2} s_i d_i, \qquad (8.224)$$

где d_i – расстояние от точки P до стороны (j, k). Тогда отношение площадей этих треугольников равно

$$\frac{S_{Pjk}}{S_{ijk}} = \frac{d_i}{h_i}.$$

Но отношение $\frac{d_i}{h_i}$, согласно определению, есть L_i. Следовательно,

$$L_i = \frac{S_{Pjk}}{S_{ijk}}. \qquad (8.225)$$

Аналогично устанавливаются значения

$$\left.\begin{aligned} L_j &= \frac{S_{Pki}}{S_{ijk}} \\ L_k &= \frac{S_{Pij}}{S_{ijk}} \end{aligned}\right\}. \qquad (8.226)$$

Поскольку

$$S_{Pjk} + S_{Pki} + S_{Pij} = S_{ijk}, \qquad (8.227)$$

постольку сумма значений естественных координат всегда равна

$$L_i + L_j + L_k = 1. \qquad (8.228)$$

Отсюда следует, что естественные координаты точки зависимы. Это вполне объяснимо, так как в двумерном случае положение точки задается всего двумя координатами.

Из (8.227) и значений естественных координат в вершинах треугольника следует, что естественные координаты при линейной интерполяции внутри треугольника играют роль функций формы

$$\left.\begin{aligned} f_i(x,y) &= L_i \\ f_j(x,y) &= L_j \\ f_k(x,y) &= L_k \end{aligned}\right\}. \qquad (8.229)$$

Следовательно, формулу для линейной интерполяции внутри треугольника можно представить как

$$F(x,y) = L_i z_i + L_j z_j + L_k z_k. \qquad (8.230)$$

Если координаты в глобальной системе координат рассматривать как функции, заданные своими значениями в вершинах треугольника, то можно написать полезные соотношения

$$\left.\begin{aligned} x &= L_i x_i + L_j x_j + L_k x_k \\ y &= L_i y_i + L_j y_j + L_k y_k \\ L_i &+ L_j + L_k = 1 \end{aligned}\right\}. \qquad (8.231)$$

Преимущества использования естественных координат проявляются, когда внутри треугольника осуществляется интерполяция второго и третьего порядков.

Точность представления топографической поверхности при прочих равных условиях тем выше, чем выше плотность исходных точек. Но съемка дополнительных точек влечет за собой увеличение затрат времени и стоимости исходных данных. Между тем точность моделирования можно заметно повысить, если вместо интерполяционных полиномов (8.210) использовать полиномы более высокого порядка. Среди таких полиномов наиболее употребительными являются полиномы второго

$$F(x,y) = \alpha_0 + \alpha_1 x + \alpha_2 y + \alpha_3 x^2 + \alpha_4 xy + \alpha_5 y^2 \qquad (8.232)$$

и третьего порядка

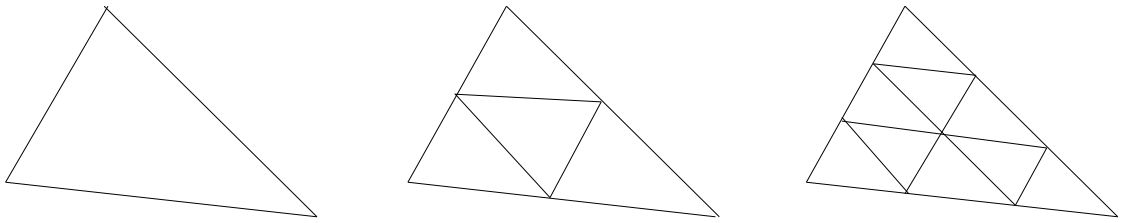
$$F(x,y) = \alpha_0 + \alpha_1 x + \alpha_2 y + \alpha_3 x^2 + \alpha_4 xy + \\ + \alpha_5 y^2 + \alpha_6 x^3 + \alpha_7 x^2 y + \alpha_8 xy^2 + \alpha_9 y.^3 \qquad (8.233)$$

В качестве базисных функций иногда также используются рациональные функции и тригонометрические полиномы.

Выше мы видели, что если в качестве базисных функций выбираются линейные полиномы, то коэффициенты полинома определяются значениями функции в вершинах треугольников. Но если степень интерполяционного полинома выше, то этих значений уже недостаточно.

Если базисные функции являются полиномами второй степени, то кроме указанных значений требуются значения функции в некоторых дополнительных точках, например, в средних точках сторон треугольника (рис. 8.76, б). При построении кубического полинома определяются 10 параметров, поэтому могут

использоваться значения функции в вершинах, две дополнительные точки на каждой его стороне и одна точка внутри треугольника (рис. 8.76, в); а также возможны другие варианты [11], [15]. Но в любом из них возникает задача определения интерполяционного полинома.



а) б) в)

Рис. 8.76. Треугольные элементы 1, 2, и 3-го порядков

Для решения этой задачи могут использоваться способы, заимствованные из метода конечных элементов, в частности – функции формы.

Общая формула для вычисления функций формы треугольных элементов любого порядка имеет вид произведения

$$f_\beta = \prod_{\gamma=1}^{n} \frac{\varphi_\gamma}{\varphi_{\gamma|L_1,\,L_2,\,L_3}}, \qquad (8.234)$$

где n – порядок треугольного элемента, φ_γ – функции естественных координат L_1, L_2 и L_3. Порядок треугольника есть величина $n = m - 1$, где m – число точек на каждой стороне.

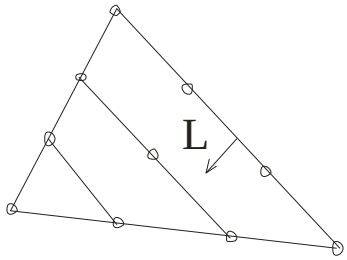


Рис. 8.77. Линии *L = const*

Функции φ_γ в числителе определяются из уравнений n линий, проходящих через все точки в треугольнике. На рис. 8.77 показаны направление отсчета одной из L-координат и линии $L = const$, проходящие через узлы треугольника. Если рассматривается прямая $L = const$, то функция в числителе φ_γ определяется выражением $\varphi_\gamma = L_1 - c$. Знаменатель представляет собой значение функции φ_γ в точке, для которой вычисляется функция формы.

На рис. 8.78 показано расположение точек на квадратичных и кубичных треугольных элементах, а в табл. 8.7 приведены выражения для функций формы в этих точках.

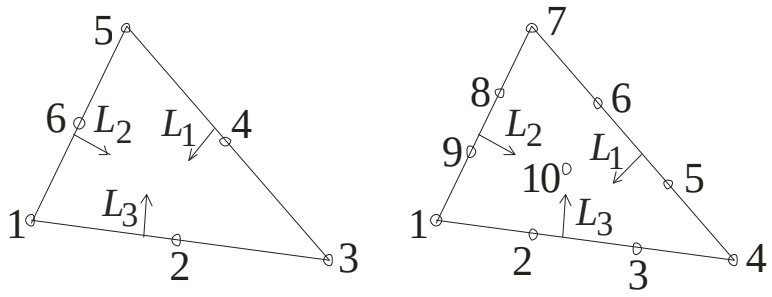


Рис. 8.78. Узлы квадратичного и кубичного элементов

Таблица 8.7. Функции формы для треугольных элементов

Квадратичный элемент	Кубичный элемент
$f_1 = L_1(2L_1 - 1)$	$f_1 = \frac{1}{2} L_1(3L_1 - 1)(3L_1 - 2)$
$f_2 = 4L_1L_2$	$f_2 = \frac{9}{2} L_1L_2(3L_1 - 1)$
$f_3 = L_2(2L_2 - 1)$	$f_3 = \frac{9}{2} L_1L_2(3L_2 - 1)$
$f_4 = 4L_2L_3$	$f_4 = \frac{1}{2} L_2(3L_2 - 1)(3L_2 - 2)$
$f_5 = L_3(2L_3 - 1)$	$f_5 = \frac{9}{2} L_2L_3(3L_2 - 1)$
$f_6 = 4L_1L_3$	$f_6 = \frac{9}{2} L_2L_3(3L_3 - 1)$
	$f_7 = \frac{1}{2} L_3(3L_3 - 1)(3L_3 - 2)$
	$f_8 = \frac{9}{2} L_1L_3(3L_3 - 1)$
	$f_9 = \frac{9}{2} L_1L_3(3L_1 - 1)$
	$f_{10} = 27L_1L_2L_3$

Важным и полезным свойством функций формы для треугольных элементов является их независимость от формы треугольника.

Выше считалось, что значения высот известны не только в вершинах треугольников, но и в дополнительных узлах. Но если в вершинах треугольников они измеряются инструментально, то в дополнительных узлах – вычисляются. В принципе, для их получения может использоваться вся мощь метода конечных элементов, но при моделировании топографических поверхностей он, вероятно, не применяется в силу своей сложности. Решение очень больших систем линейных уравнений (порядка 10^5 и больше) требует как оперативной памяти, так и процессорного времени. Поэтому МКЭ находит применение прежде всего либо там, где требуется высокая точность (например,

при расчете строительных конструкций), либо там, где задачи имеют существенно меньшую размерность.

Поэтому для вычисления высот в дополнительных узлах могут использоваться методы, не столь теоретически обоснованные, но практичные. Во-первых, высоты в точках на сторонах треугольников должны вычисляться как точки сгущения структурных линий. Во-вторых, на ограниченном числе точек может строиться гладкая поверхность, и по ней могут вычисляться высоты в нужных точках.

На рис. 8.79 представлен фрагмент триангуляции. Пусть требуется определить высоты в помеченных узлах ребра (i, j). Тогда можно выбрать все вершины, смежные вершинам i и j, на полученном множестве точек построить гладкую поверхность и вычислить по ней значения высот в требуемых узлах. Если требуется вычислять значения во внутренних узлах треугольника (i, j, k), то для построения куска гладкой поверхности могут привлекаться все вершины, смежные с вершинами i, j, k.

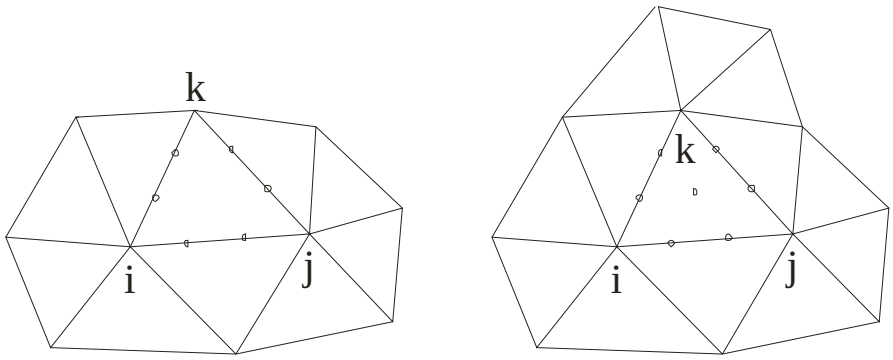
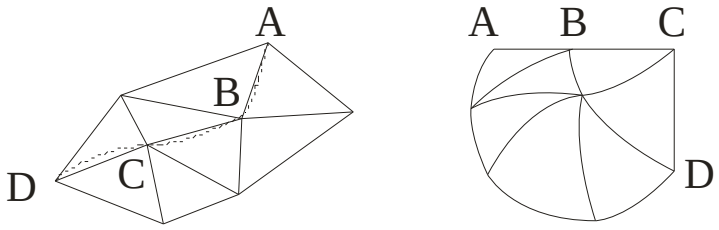


Рис. 8.79. К вычислению высот в узлах

В принципе, построение такой гладкой поверхности может осуществляться разными способами, но желательно, чтобы используемые функции по своему виду были однородными. Такими функциями являются, в частности, мультиквадрики. В дальнейшем, после вычисления значений во всех дополнительных узлах, эти функции не используются, а применяются описанные выше интерполяционные полиномы. Преимуществом такого подхода является то, что не приходится решать системы линейных уравнений очень высокого порядка, и триангуляция может обрабатываться блоками на компьютерах с малой оперативной памятью.

При оценке вычислительной сложности решения задач на нерегулярной сетке треугольников неявно постулируется, что треугольники являются прямолинейными. Однако можно привести примеры, когда соединение вершин только прямыми не позволяет правильно отобразить поверхность. В таких случаях требуется построение всех треугольников или части из них с криволинейными сторонами. Если на рис. 8.80, *а* через точки *A*, *B*, *C* и *D*, принадлежащие водотоку, провести ломаную, то его изображение может выглядеть неестественным. Но если через эти же точки провести гладкую кривую, а стороны треугольников на топографической поверхности принять

прямыми, то может произойти рассогласование изображения рельефа (горизонталей) с линией тальвега, и оно может быть заметным.



а) б)

Рис. 8.80. Криволинейные треугольники

На рис. 8.80, б приведен пример несколько иной: точки A, B, C и D принадлежат границе некоторого искусственного сооружения с прямолинейными границами. Тогда стороны AB, BC, и CD должны быть отрезками прямых, а остальные стороны треугольников при необходимости могут быть криволинейными. Однако в программных комплексах криволинейные треугольники используются редко. Поэтому, чтобы обеспечить приемлемое качество моделирования и картографического отображения ситуации и топографической поверхности, исходные точки набираются с высокой плотностью. При использовании картометрического и фотограмметрического методов сбора это не столь большая проблема, и происходит всего лишь некоторое снижение производительности труда исполнителей. Указать точку на экране монитора с помощью курсора и встать с отражателем или приемником GPS на точку на местности – по затратам времени далеко не одно и то же. Поэтому при сборе данных топометрическим методом (с использованием обычных и электронных тахеометров или GPS) может происходить катастрофическое возрастание стоимости получения исходных данных.

В итоге приходится признать необходимость использования криволинейных треугольников и хранения признаков сторон треугольников, что ведет к усложнению алгоритмов обработки данных, увеличению необходимой памяти и времени обработки. Использование криволинейных треугольников и криволинейных границ контуров ведет также к усложнению решения описанной выше задачи определения принадлежности точки треугольнику или многоугольнику. Но эти и другие затраты на разработку программного обеспечения с избытком могут быть компенсированы увеличением производительности труда и повышением точности создаваемых геоинформационных моделей, а также получаемых на их основе традиционных карт и планов. Пока что разработчики программного обеспечения, как это часто случается, предпочитают перекладывать собственные проблемы на плечи пользователей.

Покрытие (а не разбиение) области моделирования элементами обладает меньшей гибкостью и иногда применяется для представления гладких поверхностей. Типичное решение заключается в том, что область моделирования разбивается на перекрывающиеся блоки прямоугольной формы (рис. 8.81). Для каждого прямоугольника в его пределах определяется индивидуальное уравнение поверхности. В полосе перекрытия положение поверхности устанавливается как среднее весовое. Поэтому ширина блоков в каждой полосе (горизонтальной или вертикальной) должна быть постоянной, но может быть различной в разных полосах. Длина блоков в пределах одной полосы может изменяться. В различных реализациях этой идеи варьируются размеры элементов, процент перекрытия, вид уравнения поверхности и вид весовых функций.

При интеграции блоков в единую модель (например, при пересчете на регулярную сетку) первоначально осуществляется сшивка блоков в каждой полосе, а затем объединение полос. Формулы для вычисления значения высоты в полосе перекрытия достаточно очевидны.

8.21. Построение плоской триангуляции

При моделировании топографических поверхностей часто используется разбиение области моделирования на непересекающиеся треугольники с вершинами в исходных точках, называемое *триангуляционным покрытием*, или *плоской триангуляцией*. Построение плоской триангуляции является обязательным первым этапом при конструировании нерегулярных кусочно-непрерывных моделей. Затем на каждом треугольном элементе строится поверхность первого или более высокого порядка. Значительная доля вычислительных затрат при этом приходится на этап построения триангуляции. В [12] эта задача даже подозревалась в принадлежности к труднорешаемым (неполиномиальным).

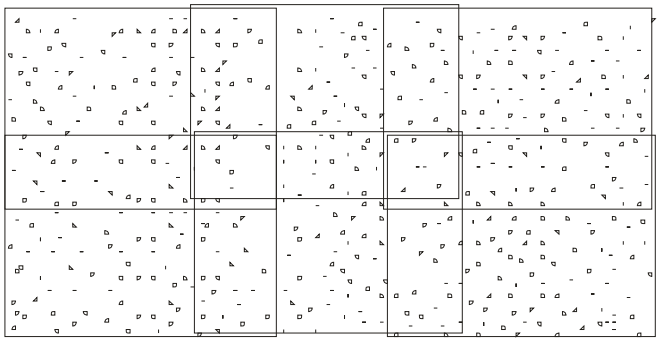
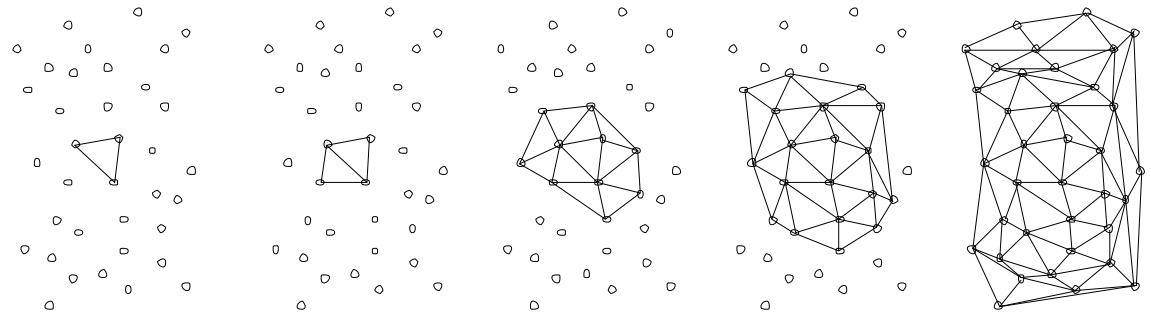


Рис. 8.81. Покрытие области

Создание сети треугольников, вершинами которых являются произвольно расположенные исходные точки, является достаточно сложной задачей, заслуживающей отдельного рассмотрения. Задача формулируется следующим образом. На заданном множестве произвольно расположенных точек $\{P_i(x_i, y_i)\}$, (i = 1, …, n) требуется построить множество треугольников $\{T_j(P_{j1}, P_{j2}, P_{j3})\}$, ($j$ = 1, …, m), отвечающих условию $T_i \bigcap T_j = \varnothing$ при $i \neq j$. Существует два общих подхода к решению поставленной задачи, в которых, так или иначе, используется граница области моделирования.

Первый подход может быть назван построением триангуляции *от центра к периферии*. При его использовании вначале выбирается некоторая точка P_j, о которой заведомо известно, что она является внутренней точкой области моделирования. Затем среди оставшихся исходных точек отыскивается точка, ближайшая к первой. Если таких точек несколько, выбирается любая из них. После этого отыскивается третья точка, сумма расстояний от которой до первых двух является минимальной. Полученные таким образом точки образуют начальный треугольник. Далее вокруг этого треугольника строится первый слой соседних треугольников, затем второй и т. д., пока граница триангуляции не совпадет с границей области моделирования (рис. 8.82). Для ускорения поиска точки могут быть предварительно отсортированы по возрастанию расстояния от первой точки.

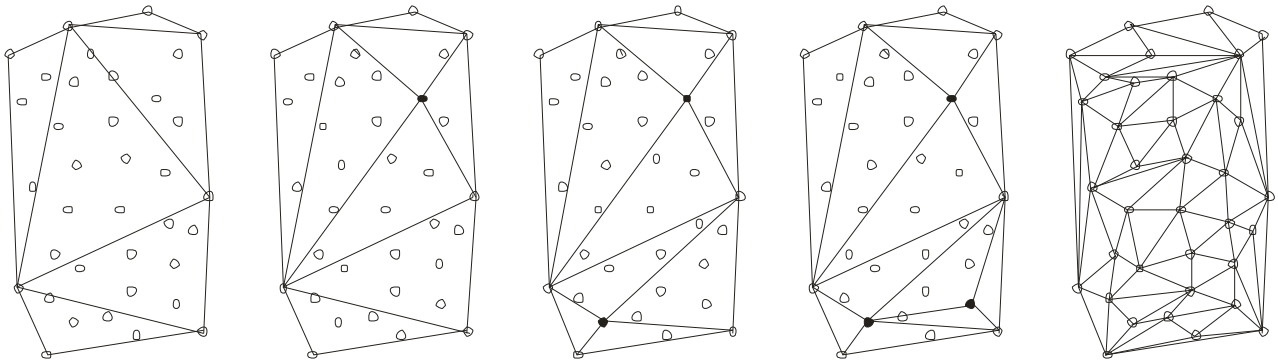


а) *б*) *в*) *г*) *д*)

Рис. 8.82. Триангуляция от центра к периферии

Различия между вариантами построения триангуляции от центра периферии сводятся к тому, остается внешняя граница триангуляции выпуклой или нет.

При втором подходе построение триангуляции осуществляется *от периферии*, то есть от границы *к центру*. Граница триангуляции часто определяется как выпуклая оболочка. *Выпуклой оболочкой* множества точек на плоскости называют выпуклый многоугольник, вершинами которого являются точки данного множества, построенный таким образом, что все остальные точки множества являются его внутренними точками. При этом возможны два варианта. В первом случае из числа точек, принадлежащих границе области моделирования, некоторым образом строится плоская триангуляция (рис. 8.83, *а*). После этого осуществляется поочередная вставка в триангуляцию оставшихся точек (рис. 8.83, *б*, *в*, *г*). Сеть треугольников при вставке каждой точки модифицируется: либо уничтожается старый треугольник и образуются три новых, либо уничтожаются два старых треугольника и создаются четыре новых (если точка попала на сторону двух смежных треугольников или близка к ней). Таким образом, данный метод основан на вставке вершин в существующую триангуляцию.

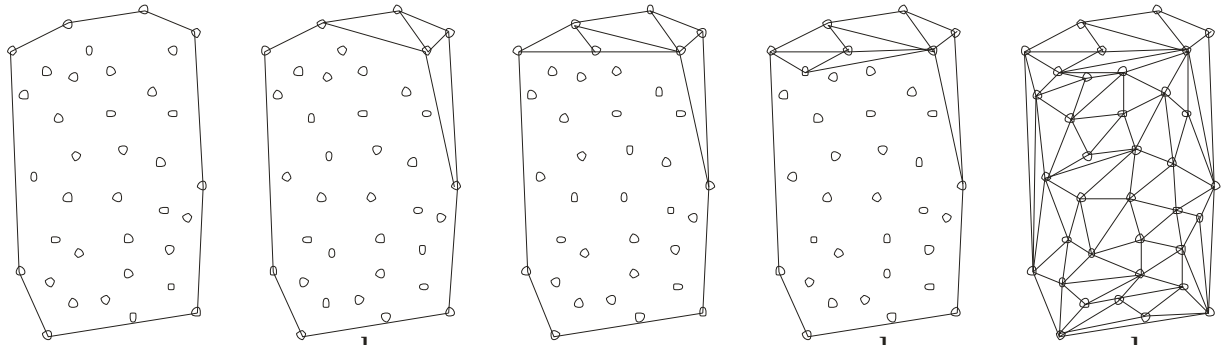


а) б) в) г) д)

Рис. 8.83. Триангуляция от периферии к центру, вариант 1

После вставки последней точки сеть будет построена, но далека от оптимальной. Поэтому приступают к ее перестройке в соответствии с некоторым критерием. В качестве критерия обычно используется минимум суммы длин сторон треугольников (оптимальная триангуляция), либо строится триангуляция Делоне – триангуляция, в которой внутрь окружности, описанной вокруг любого треугольника, не попадает никакая другая вершина триангуляции. Возможно, что следует минимизировать сумму квадратов длин сторон, так как отпадает необходимость извлечения квадратных корней, и треугольники имеют более правильную форму. Наиболее полное описание триангуляции Делоне дано в [41], где представлены пять вариантов структуры данных и 28 алгоритмов построения.

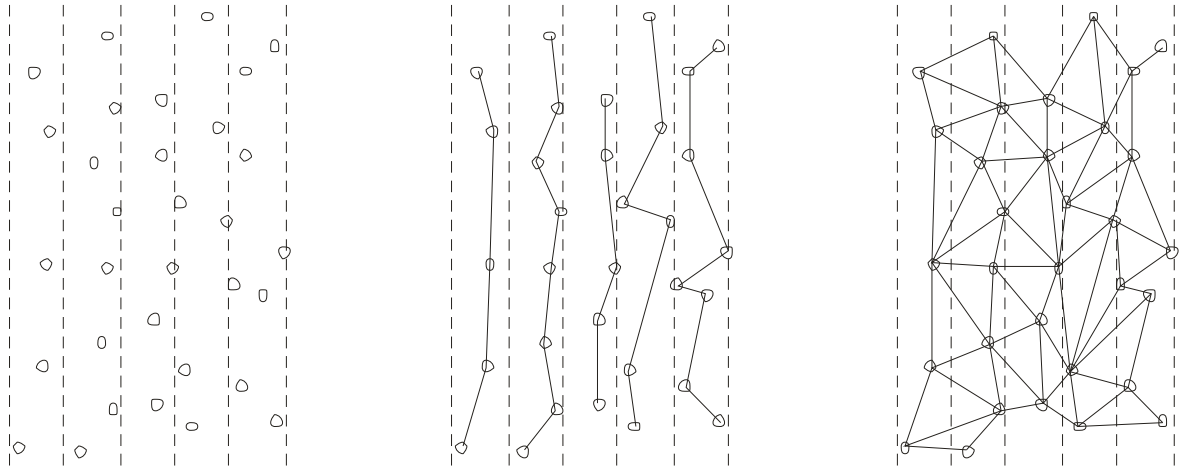
Во втором варианте построение триангуляции осуществляется наращиванием треугольников от границы области моделирования к ее середине. В процессе построения триангуляции граница области, не покрытой треугольниками, постоянно корректируется (рис. 8.84).



а) *б*) *в*) *г*) *д*)

Рис. 8.84. Триангуляция от периферии к центру, вариант 2

Еще одним способом построения плоской триангуляции является разбиение области моделирования на полосы (рис. 8.85). После этого в каждой полосе точки упорядочиваются вдоль полосы, и через полученную последовательность точек проводится ломаная. Затем выполняется соединение точек разных полос.



а) б) в)

Рис. 8.85. Построение триангуляции разбиением на полосы

Перечисленные алгоритмы характеризуются:

– высоким уровнем логической сложности – большим числом различных условий (вершины треугольника не должны лежать на одной прямой, внутри каждого треугольника не должно быть точек, треугольники не должны пересекаться и т. д.);

– низким быстродействием – большим общим числом операций по перебору точек и треугольников.

В других, более быстрых, алгоритмах используется свойство частичной упорядоченности исходных точек: их расположение по профилям (или галсам) и т. п. Но при произвольном расположении точек эти алгоритмы не могут быть применены.

В процессе построения триангуляции от периферии к центру способом вставки (вариант 1) необходимо отыскивать треугольник, в который попадает вставляемая точка. Эта же задача возникает при определении высоты в точке с координатами (x, y) по нерегулярной дискретной модели. Вначале требуется найти нужный треугольник, а уже затем определить высоту тем или иным способом.

В более общей постановке эта задача формулируется как определение принадлежности точки произвольному многоугольнику. Существует два основных способа определения принадлежности или непринадлежности некоторой точки заданному многоугольнику (рис. 8.86). Для решения указанной задачи первым способом через эту точку проводится произвольная прямая, но проще такую прямую провести параллельно одной из осей координат. Если на данной прямой слева от определяемой точки число точек пересечений прямой со сторонами многоугольника нечетно, то определяемая точка находится внутри многоугольника, если число таких точек четно или равно нулю, то определяемая точка находится вне многоугольника (рис. 8.86, а, в) .

Решение задачи определения принадлежности точки многоугольнику вторым способом состоит в последовательном суммировании углов, образованных двумя соседними направлениями с определяемой точки на

вершины многоугольника. При этом углы, определяемые, например, по ходу часовой стрелки, считаются положительными, а откладываемые в противоположном направлении – отрицательными. Если сумма всех таких углов

$$S = (\alpha_{O2} - \alpha_{O1}) + ... + (\alpha_{On} - \alpha_{O1}),$$

где α_{Oi} – направление (азимут) с точки O на точку i, равна нулю, то определяемая точка лежит вне многоугольника, если же сумма равна 360°, то точка находится внутри многоугольника. Для случая треугольника (рис. 8.86, а) это условие легко проверить. Для произвольного многоугольника (рис. 8.86, г) данное условие также справедливо. На рис. 8.86 точка A принадлежит многоугольнику, а точка B – нет.

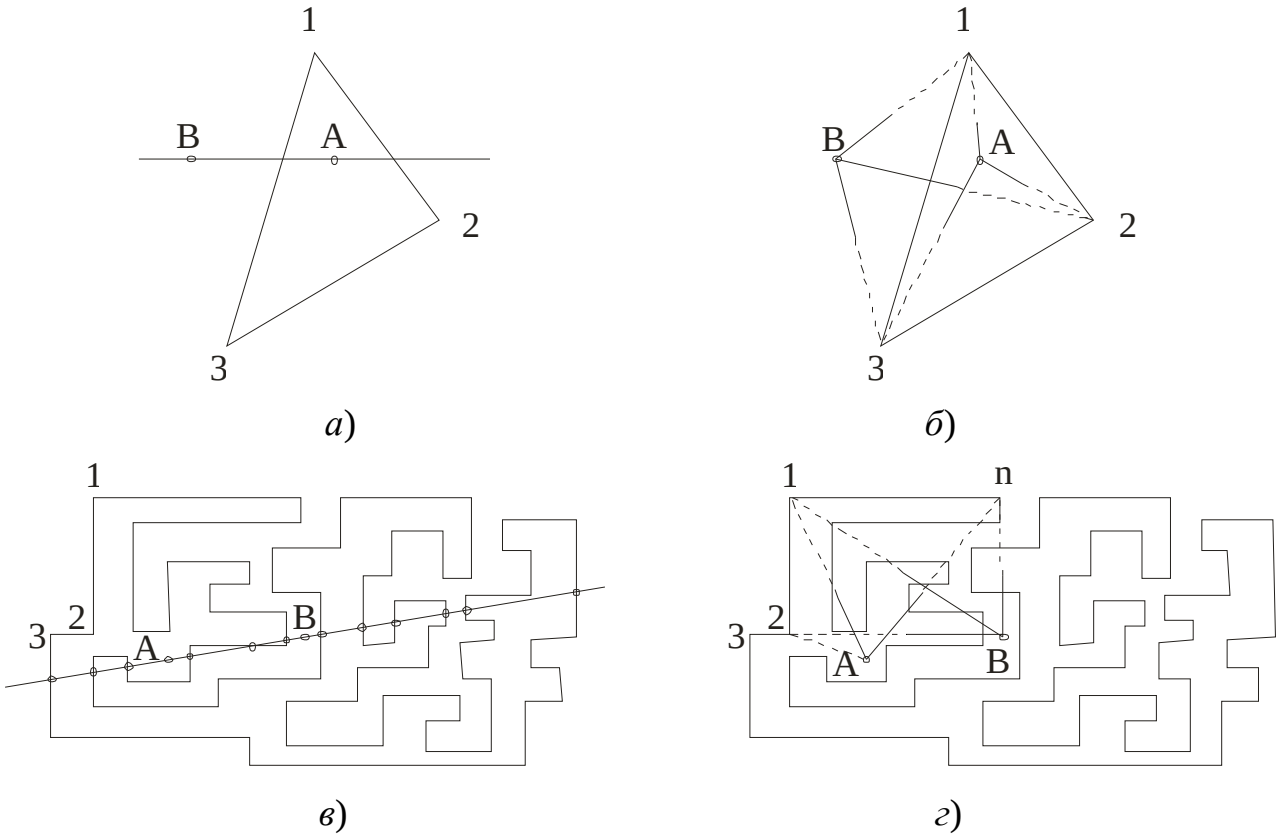


Рис. 8.86. Принадлежность многоугольнику

8.22. Волновые алгоритмы построения плоской триангуляции

Недостатком рассмотренных выше алгоритмов построения триангуляции является необходимость предварительного определения границы области моделирования. Она может указываться пользователем в интерактивном режиме либо определяться программным путем как выпуклая оболочка множества исходных точек. В первом случае граница может быть произвольным многоугольником, но требуются ощутимые затраты ручного труда. Во втором случае увеличиваются затраты процессорного времени. Кроме того, граница имеет вид выпуклого многоугольника, появляются лишние ребра и треугольники неудовлетворительной формы вблизи границы области, и требуется их удаление и корректировка границы.

К недостаткам рассмотренных методов можно отнести также то, что участок моделирования должен быть связной областью. Несвязные области должны моделироваться раздельно. При моделировании многосвязных областей (областей с «дырами») может потребоваться удаление значительного числа треугольников, построенных программным путем.

Ниже рассматриваются два метода построения плоской триангуляции, принципиально отличающиеся от рассмотренных и основанные на дискретизации области моделирования. Первый из них может быть назван прямым волновым алгоритмом. Его описание здесь приводится для объяснения второго, более эффективного алгоритма, названного обратным волновым алгоритмом.

Пусть, как и прежде, задано множество точек, произвольно расположенных в области моделирования, границы которой априори не известны. Построение плоской триангуляции с помощью *прямого волнового алгоритма* выполняется следующим образом.

Строится сетка квадратов с достаточно малым интервалом между ее узлами, так чтобы все исходные точки попадали внутрь этой сетки (рис. 8.87, *а*). Вершинам сетки в оперативной памяти соответствует матрица H_{mn}, где m и n – число узлов по осям X и Y соответственно. Этим мы заменяем область моделирования, являющуюся всюду плотным множеством, множеством дискретных точек – матрицей H_{mn}.

Прямой и обратный алгоритмы разбиваются на два этапа:

1. определение значений элементов матрицы H;
2. анализ матрицы H.

В обоих алгоритмах матрица H предварительно обнуляется.

На первом этапе прямого алгоритма для каждого узла квадратной сетки (элемента матрицы H) методом итераций определяется ближайшая исходная точка и ее номер присваивается узлу. С этой целью в первой итерации последовательно просматриваются все исходные точки. Для каждой точки определяется квадрат, в который она попадает, и его вершинам присваивается значение номера этой точки (рис. 8.87, *б*).

Вновь повторяется цикл по числу исходных точек, при этом порядковый номер точки присваивается не вершинам соответствующих квадратов, а соседним с ними узлам сетки, если значения номеров ближайших точек в них равны нулю (рис. 8.87, *в*). При следующем цикле по исходным точкам присваиваются значения следующему ряду узлов вокруг каждой исходной точки и т. д., пока не останется узлов с нулевыми значениями. В результате получим картину, представленную на рис. 8.87, *г* и рис. 8.88, на котором показаны исходные точки, сетка квадратов и (будущая) триангуляция.

Процесс присваивания элементам матрицы H номеров ближайших к ним исходных точек напоминает распространение волн, источником которых являются исходные точки, почему алгоритм и назван волновым (см. рис. 8.87). Своеобразие этих волн в том, что они имеют форму не окружностей, а квадратов. Как только волна от точки с номером i достигает некоторого узла,

который не был достигнут ранее другой волной (узел в таком случае имеет первоначальное значение 0), ему присваивается значение *i*. Если узел уже был достигнут другой волной (тогда элемент матрицы *H* имеет ненулевое значение), то сохраняется прежнее значение и в этом направлении волна в дальнейшем не распространяется. Происходит как бы взаимное гашение встретившихся волн.

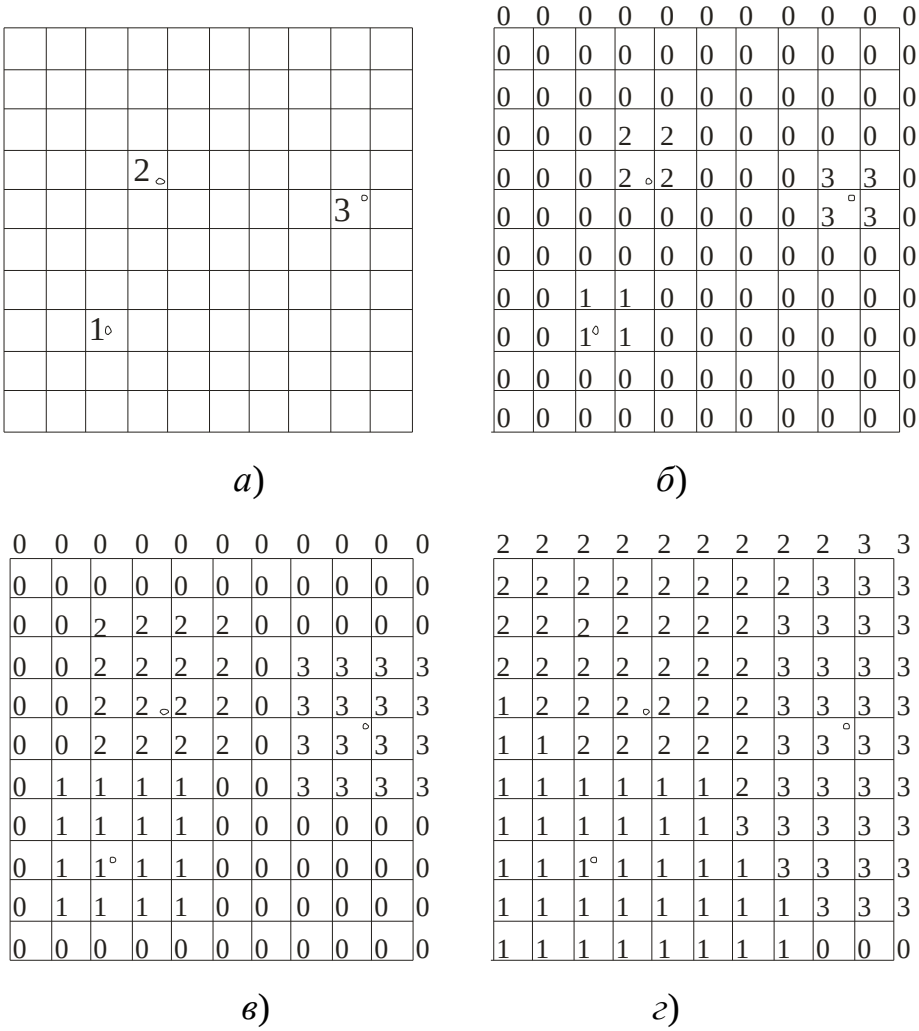


Рис. 8.87. Выполнение итераций в прямом алгоритме

Если трем вершинам некоторого квадрата присвоены номера различных между собой исходных точек (на рис. 8.88 они обозначены знаком «+»), то эти точки образуют треугольник.

Если четырем вершинам некоторого квадрата присвоены различные между собой значения номеров исходных точек, то необходимо сделать выбор одной из двух пар возможных треугольников (рис. 8.89). Пусть, например, вершинам квадрата присвоены номера исходных точек *A*, *B*, *C*, *D* (обход вершин квадрата совершается по часовой стрелке или против часовой стрелки). Если расстояние $(AC)^2 < (BD)^2$, то выбираются треугольники *ABC* и *ACD*, в противном случае – треугольники *ABD* и *BCD*.

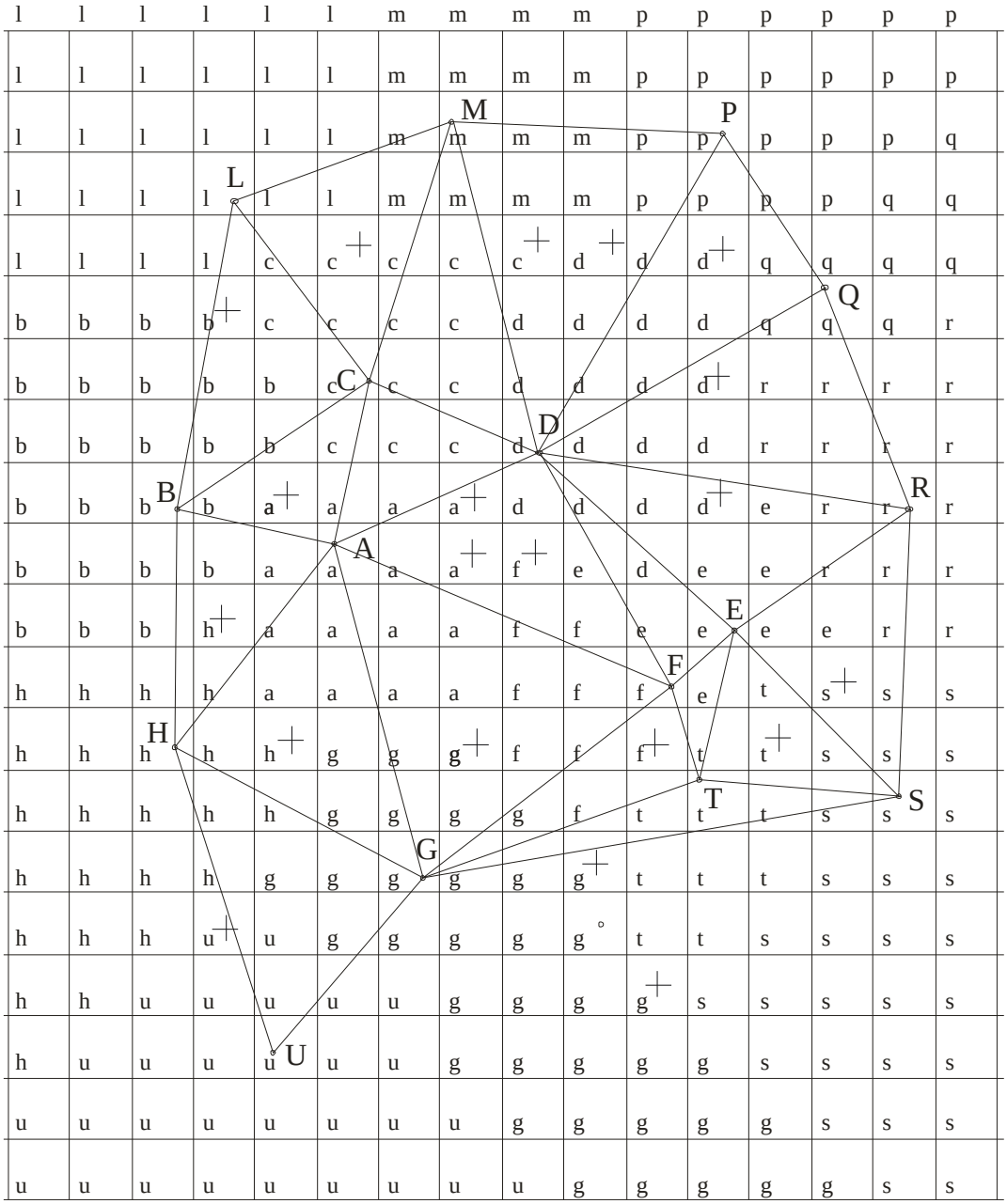


Рис. 8.88. Построение плоской триангуляции

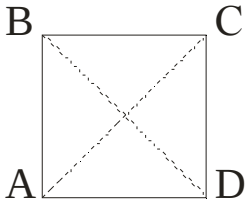


Рис. 8.89. Выбор треугольников

Важное свойство полученной таким образом матрицы *H* состоит в том, что каждый треугольник встречается в ней только один раз, и треугольники являются непересекающимися. Отсюда следует второй этап построения плоской триангуляции.

На втором этапе построения плоской триангуляции последовательно просматриваются все квадраты сетки, и если в каком-либо из них есть хотя бы три различных номера исходных точек, то эти номера пересылаются в список треугольников. По завершении просмотра всех квадратов задача будет решена – получен список треугольников. Граница области моделирования (не обязательно выпуклая) может быть получена как совокупность сторон, каждая из которых входит только в один треугольник.

В некоторых случаях, например, когда область моделирования имеет подковообразную форму, для построения сетки треугольников могут потребоваться значительные затраты процессорного времени и возникнуть треугольники неудовлетворительной формы (например, треугольник GST на рис. 8.88). Поэтому на первом этапе есть смысл ограничить число циклов по исходным точкам так, чтобы максимальное удаление просматриваемых узлов квадратной сетки от исходных точек не превышало некоторого заданного допуска, скажем, 2/3 наибольшего допускаемого инструкциями расстояния между точками. Ограничение просматриваемых узлов их удалением от исходных точек дает возможность:

– существенно снизить вычислительные затраты;

– представлять участки моделирования, являющиеся односвязными, многосвязными или несвязными областями (рис. 8.91);

– избавиться от появления лишних треугольников;

– более точно представлять границы области моделирования.

Обратный волновой алгоритм отличается от прямого только направлением движения волны. В прямом алгоритме волны распространяются от каждой исходной точки к периферии, в обратном – от периферии к точкам (рис. 8.90). На рис. 8.90, *б* и 8.90, *в* следует обратить внимание на то, как последующие значения номеров ближайших точек заменяют предыдущие значения. Это дает возможность избавиться от проверки того, что узел сетки квадратов уже был достигнут другой волной. В результате уменьшаются сложность алгоритма и время построения триангуляции.

Время обработки на первом этапе в обоих алгоритмах можно сократить, если перед вычислением матрицы H для каждой исходной точки по ее координатам определить и сохранить в памяти индексы квадрата, в который попадает точка. Тогда мы избавимся от необходимости вычислять индексы в каждой итерации, выполнив эту работу только один раз.

Кроме того, имеется еще одна возможность для ускорения алгоритма при одновременном снижении его сложности. Нетрудно понять, что на первом этапе перед присвоением узлу сетки квадратов номера ближайшей исходной точки мы должны выполнять проверку существования такого узла. Но от такой проверки можно избавиться, если сетку квадратов создать с некоторым перекрытием. Для этого сетка квадратов должна покрывать не область, ограниченную прямыми $x_{\min} - s/2$, $x_{\max} + s/2$, $y_{\min} - s/2$ и $y_{\max} + s/2$, а область, ограниченную прямыми $x_{\min} - k$, $x_{\max} + k$, $y_{\min} - k$ и $y_{\max} + k$, где $k = d/s$ и s – сторона квадрата, а d – максимальное допускаемое инструкциями расстояние между исходными точками. Для нас d – это расстояние, дальше которого просмотр узлов не производится.

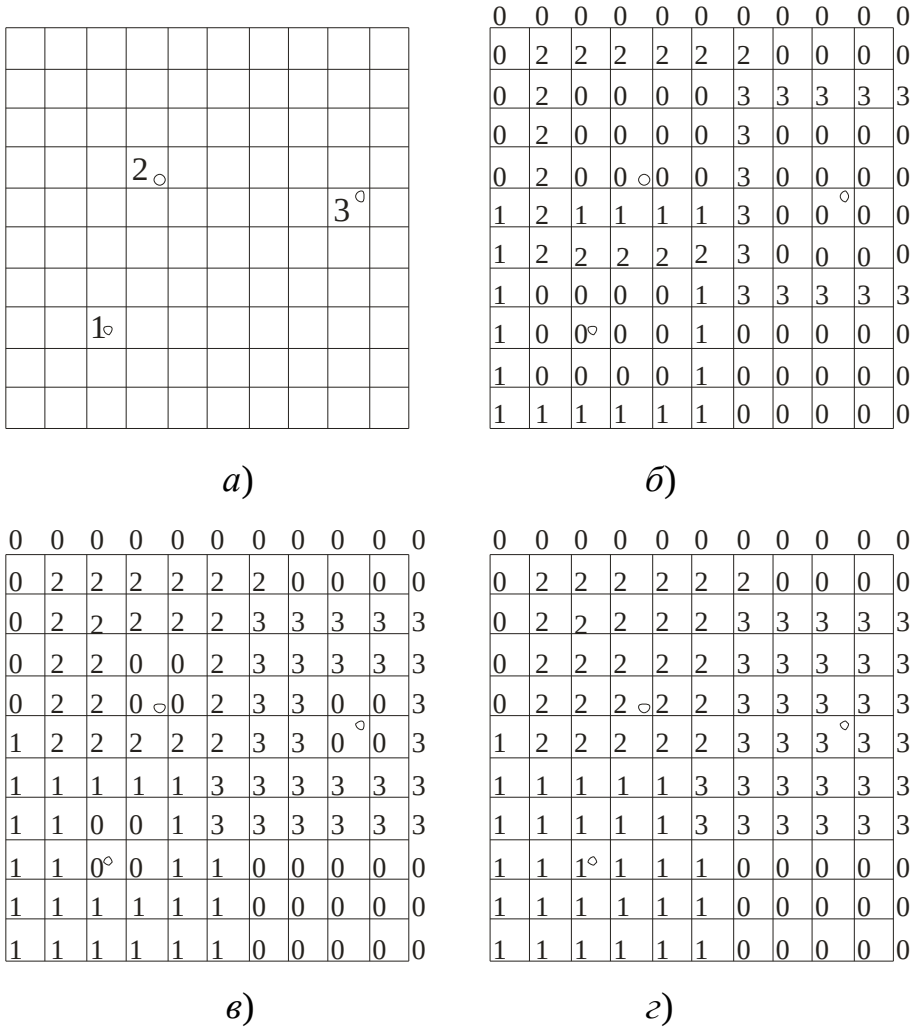


Рис. 8.90. Выполнение итераций в обратном алгоритме

В заключение рассмотрим преимущества и недостатки волновых алгоритмов построения плоской триангуляции, сравнив их с ранее известными методами.

1. Волновые алгоритмы дают возможность построения сетки треугольников для областей моделирования сложной конфигурации (рис. 8.91) как единого целого. Использование других алгоритмов потребовало бы выделения и раздельной обработки областей *A*, *B*, *C*. Участок моделирования может быть связной или несвязной, одно- или многосвязной областью. Так, на рис. 8.91 весь участок моделирования является несвязной областью, которая представляет собой совокупность связных областей *A*, *B* и *C*. Области *A* и *B* являются односвязными, а область *C* – многосвязной. Связные области могут быть выпуклыми, как область *A*, либо невыпуклыми, как области *B* и *C*. Разумеется, минимальное

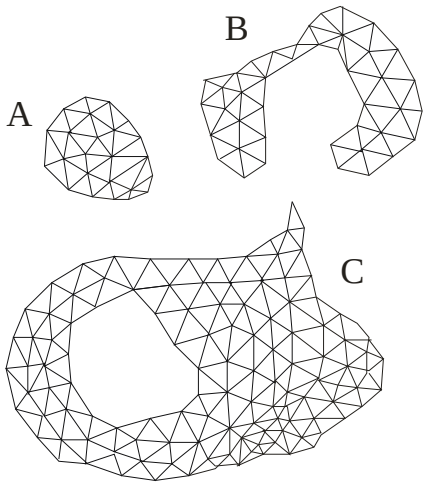


Рис. 8.91. Область сложной конфигурации

расстояние между двумя областями должно быть несколько больше допустимого расстояния между исходными точками.

Если бы плоская триангуляция строилась с помощью известных ранее методов, то потребовалось бы предварительно описывать границы каждой связной области моделирования либо строить их программным путем. Однако, программное построение границы области привело бы к покрытию всей области моделирования одной сплошной сетью триангуляции, и потребовалось бы перечисление большого числа удаляемых треугольников.

2. Еще одним достоинством волновых алгоритмов является их высокое быстродействие. Вычислительная сложность обоих алгоритмов характеризуется линейной зависимостью от числа исходных точек и числа узлов сетки квадратов:

$$O_n = O_{1n}(k) + O_2(mn);$$

$$O_{\text{о}} = O_{1\text{о}}(k) + O_2(mn),$$

где O_n – вычислительная сложность прямого, а $O_{\text{о}}$ – обратного алгоритма; O_{1n} и $O_{1\text{о}}$ – вычислительная сложность первого этапа соответственно для прямого и обратного алгоритмов; O_2 – вычислительная сложность второго этапа; k – число исходных точек; mn – число узлов квадратной сетки. Поскольку обратный волновой алгоритм более эффективен:

$$O_{1n}(k) > O_{1\text{о}}(k),$$

программная реализация прямого алгоритма становится нецелесообразной.

Вычислительные затраты для волновых алгоритмов столь малы, что они могут использоваться даже для отбраковки ошибок в исходных данных.

3. Недостатком волновых алгоритмов является потребность в дополнительной памяти для хранения сетки квадратов. Однако этот недостаток может быть сглажен перекрытием массивов для квадратной и треугольной сетки и размещением первого элемента матрицы H в оперативной памяти по тому же адресу, что и треугольник с номером $i >> 1$. Возможна также запись треугольников сразу во внешнюю память.

4. При использовании волновых алгоритмов могут возникнуть определенные сложности при попадании нескольких исходных точек в один квадрат сетки квадратов. Эта проблема может решаться как уменьшением размеров квадрата, так и последующей модификацией сетки треугольников, то есть вставкой вершин в триангуляцию, что более реально.

Чтобы сократить время поиска треугольника, в который попадает вставляемая исходная точка, у каждой исходной точки можно дополнительно создать указатель на другую исходную точку, попадающую в тот же квадрат. Тогда точки, попадающие в один квадрат, будут представлять собой линейный список, который можно сделать кольцевым.

5. С точки зрения «правильности» построенной триангуляции волновые алгоритмы, как и любые другие методы, не являются абсолютно безукоризненными. Видимо, построить безошибочно сетку треугольников программным путем в принципе невозможно, так как из того факта, что

расстояние между двумя точками мало, не обязательно следует, что они образуют сторону треугольника. Эта зависимость носит лишь вероятностный характер.

Чтобы убедиться в этом, достаточно рассмотреть частный случай (рис. 8.92). Пусть точки 1 и 3 принадлежат тальвегу или водоразделу, а 2 и 4 – некоторые характерные точки на склонах. Тогда, как и в ряде других случаев, которые можно называть *аномалиями триангуляции*, применение формального критерия минимума суммы сторон не обеспечивает правильного построения триангуляции.

Рис. 8.92. Пример аномалии

6. По этой причине часто используется так называемая *триангуляция с ограничениями*. Такими ограничениями являются структурные линии топографической поверхности: тальвеги, водоразделы, линии изменения кривизны склонов, бровки оврагов и т. п. Ограничения заключаются в том, что структурные линии не должны пересекаться сторонами треугольников. При реализации других алгоритмов построения триангуляции учет структурных линий может привести к возрастанию логической сложности алгоритмов.

Использование волновых алгоритмов позволяет легко включать структурные линии в обработку. Для этого достаточно выполнить сгущение структурных линий так, чтобы расстояние между двумя точками на структурной линии находилось в диапазоне от $1{,}5\,s$ до $2\,s$, где s – длина стороны квадрата. Сгущение точек структурных линий – это задача интерполяции заданной таблично функции одной переменной, которая намного проще аналогичной задачи для функции двух переменных. Такая интерполяция структурных линий может быть гладкой по высоте и/или в плане, что позволит повысить точность моделирования поверхности в целом.

Полученные в результате интерполяции структурных линий дополнительные точки включаются в список исходных точек. Далее без каких-либо изменений может использоваться любой из описанных волновых алгоритмов.

7. Правильно построенная сетка треугольников в действительности не обладает тем свойством, что сумма длин ее сторон (триангуляция Делоне) или их квадратов минимальна. Критерий минимума является формальным; его использование обеспечивает лишь хорошее приближение к действительности. При построении плоской триангуляции с помощью волновых алгоритмов явно не предполагается достижение некоторого минимума. Если предположить, что волны являются световыми, и учесть, что свет распространяется по кратчайшему пути, то интуитивно ясно, что полученное решение является субоптимальным. Немаловажным его достоинством является незначительное по сравнению с другими методами число корректировок первоначальной триангуляции.

8. Волновые алгоритмы построения триангуляции, особенно – обратный алгоритм, отличаются невысокой логической сложностью. Их программная реализация не вызывает никаких проблем.

9. Главное достоинство волновых алгоритмов, которое в настоящее время (на машинах с фоннеймановской архитектурой) не может быть использовано, – возможность чрезвычайно высокого распараллеливания вычислений. Распространение волн от всех исходных точек, как и анализ всех квадратов, может осуществляться параллельно. Данное свойство волновых алгоритмов станет решающим, когда массовое распространение получат матричные процессоры или нечто им подобное, то есть машины с действительно параллельной обработкой, а не ее имитацией на конвейерных процессорах.

Прямой и обратный волновые алгоритмы могут быть модифицированы: вместо сетки квадратов может использоваться сетка равносторонних треугольников (рис. 8.92, *а*). Необходимость создания треугольника нерегулярной сетки возникает при этом тогда, когда трем вершинам некоторого правильного треугольника присвоены номера различных исходных точек. На рис. 8.92, *а* такой треугольник отмечен знаком «+». Однако целесообразность такой модификации под вопросом: возможно, что в некоторых случаях треугольники будут иметь более правильную форму, но оба алгоритма существенно усложнятся.

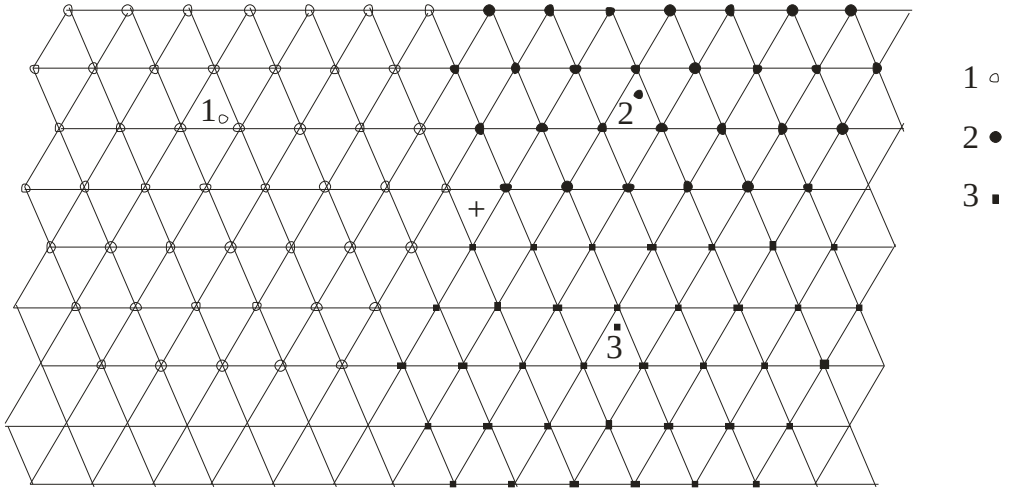


Рис. 8.92, *а*. Волновой алгоритм на сетке треугольников

В модифицированном виде волновой алгоритм использовался в разработанном в МГУ пакете программ «МАГ» (Моделирование и Анализ в Геонауках) [39], [40].

8.23. Неявная триангуляция

Рассмотренные выше модели топографической поверхности на нерегулярной сетке треугольников являются статическими, раз и навсегда созданными. Но при выводе картографического изображения на экран монитора его детальность должна зависеть от масштаба. Нет необходимости и возможности изображать в мелком масштабе все детали ситуации и рельефа, поскольку такое изображение становится совершенно нечитаемым. Таким

образом, мы приходим к известной *задаче генерализации*. Эта задача возникает также при получении карт по картам более крупных масштабов.

Генерализацию ситуации мы здесь рассматривать не будем, а ограничимся только генерализацией нерегулярных моделей топографических поверхностей. Тогда эта задача сводится к задаче исключения из существующей триангуляции тех или иных вершин и ее перестроению. Заметим, что подобная задача должна решаться в режиме реального времени. Психологами экспериментально было установлено, что максимальное время ожидания вывода изображения на экран не должно превышать четырех секунд. При большем времени ожидания пользователи испытывают сильное раздражение и могут отказаться от применения программы.

Чтобы повысить скорость вывода изображения модели на экран, с каждой вершиной триангуляции предварительно может быть связан признак, указывающий необходимость ее учета при построении изображения в требуемом масштабе. С этой целью каждой вершине v_i можно поставить в соответствие целое число m_i и принять соглашение о включении ее в триангуляцию, если выполняется соотношение $m_i \le M$, где M – знаменатель масштаба изображения.

Очевидно, что вычисление m_i для каждой из тысяч или десятков тысяч точек (на каждом номенклатурном листе топографической карты) должно осуществляться программным путем, поскольку решение данной задачи в интерактивном режиме если и возможно, то крайне неэффективно.

Вычисление признака m_i основывается на понятиях горизонтальной и вертикальной генерализации. Критерий *горизонтальной генерализации* определяет минимальные размеры объекта (в данном случае – треугольника), подлежащего отображению в заданном масштабе. Если размеры треугольника, например, его площадь или наибольшая сторона, меньше заданного, то он не должен отображаться. Но просто исключить треугольник нельзя, поскольку он связан с соседними.

Возможен следующий способ исключения вершин из триангуляции по критерию горизонтальной генерализации: если площадь каждого треугольника, инцидентного вершине v_i , меньше допуска для данного масштаба, то такая вершина исключается и триангуляция перестраивается (рис. 8.93). Поскольку часть инцидентных вершине треугольников может иметь площадь как больше, так и меньше допуска, то при решении данной задачи может использоваться аппарат теории размытых множеств.

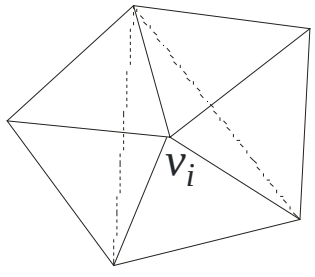


Рис. 8.93. Исключение вершины

В соответствии с критерием *вертикальной генерализации*, точка должна исключаться из триангуляции, если ее отклонение от плоскости меньше допуска h для данного масштаба. Пусть требуется определить, подлежит ли вершина v_i

исключению из триангуляции (см. рис. 8.93). Для решения этой задачи необходимо проанализировать подграф триангуляции, включающий эту вершину и все вершины, смежные с ней. Выделенный фрагмент триангуляции перестраивается в предположении, что вершина подлежит исключению из триангуляции. После этого определяется новый треугольник, в который попадает вершина v_i и вычисляется ее отклонение от плоскости этого треугольника. Если отклонение меньше допуска h, то она действительно подлежит исключению из триангуляции по критерию вертикальной генерализации.

Задача определения целесообразности или нецелесообразности включения конкретной вершины в триангуляцию, в общем-то, является частным вопросом. Более важной является проблема общей организации алгоритма генерализации нерегулярной модели топографической поверхности. Эта проблема может быть даже названа стратегической (для этой задачи). Известны два принципиально различающихся метода ее решения.

Первый метод состоит в том, что первоначально в триангуляцию включаются все вершины, отображаемые в самом крупном масштабе. Затем по мере уменьшения масштаба изображения вершины триангуляции последовательно удаляются, то есть им приписывается определенное значение признака отображения m_i, и триангуляция пересчитывается.

Второй метод решения задачи генерализации нерегулярной модели топографической поверхности начинается с самого мелкого масштаба, когда в триангуляцию включаются самые характерные точки. В принципе, начальная триангуляция при таком подходе может состоять всего из нескольких треугольников, построенных на граничных вершинах триангуляции, в пределе – из двух или даже из одного треугольника. Затем оставшиеся вершины последовательно анализируются, и триангуляция дополняется самыми характерными из них и перестраивается для каждого масштаба.

Данный метод представляет, по существу, решение обратной задачи генерализации модели топографической поверхности – все большее уточнение модели. Возражений такой подход вызывать не может, поскольку в обоих случаях мы получаем значение признака вывода для каждой вершины триангуляции. Важно также то, что с вычислительной точки зрения данный метод считается более эффективным, чем первый.

После того как для каждой точки триангуляции определено значение признака m_i, вершины (или их индексы) могут быть упорядочены по значению этого признака в убывающем порядке. Тогда при выводе изображения на экран не будет необходимости просматривать все вершины, а достаточно будет проверки с начала массива до первой точки, у которой $m_i < M$. Кроме того, в первую очередь необходимо выполнять упорядочивание по расположению точек (с использованием квадродеревьев или клеток), а уже затем по признаку масштаба.

Таким образом, рассмотренный метод представления плоской триангуляции может быть назван *динамическим*, или *динамической*

триангуляцией. Но, отдавая должное его авторам из Кембриджского университета, будем использовать предложенный ими термин – *неявная* (implicit) *триангуляция* [37].

Другая причина желательности использования неявной триангуляции – необходимость согласования изображения топографической поверхности с изображением ситуации. Данной проблеме уделяется недостаточно внимания. Как правило, модели ситуации и модели топографической поверхности являются статическими, рассчитанными на вполне определенный масштаб издаваемых карт. Модели обоих типов при этом более или менее согласованы друг с другом. Если есть какие-то несогласованности, то они устраняются на картографическом изображении в интерактивном режиме при подготовке карт к изданию. Модели же (базы данных) при этом остаются неизменными.

Источником различных недоразумений и дополнительных ошибок при моделировании является раздельное представление модели ситуации и модели топографической поверхности на одну и ту же территорию. Они рассматриваются как самостоятельные сущности, что позволяет разработчикам до поры до времени бороться со сложностью задачи представления данных о земной поверхности, откладывая ее полное решение на более поздние сроки. Но если проблеме согласования данных о топографической поверхности с данными о ситуации не уделять должного внимания, то создаваемые модели никогда не будут корректными.

Примерами объектов ситуации, которые необходимо учитывать при создании моделей топографических поверхностей, могут служить овраги, дамбы, насыпи и выемки на дорогах, карьеры и т. п. Овраги иногда снабжаются подписью максимальной глубины, хотя в действительности глубина оврага является переменной величиной. Поэтому профиль, пересекающий овраг и построенный программным путем, может содержать значительные ошибки, даже если овраг и был включен в обработку. В худшем случае овраг может быть вообще не учтен. Программный расчет зоны затопления при паводках может давать совершенно неадекватную картину, если при этом не учитываются дамбы. Подобные примеры можно продолжать, но и приведенных достаточно для понимания важности использования данных о ситуации при моделировании топографических поверхностей.

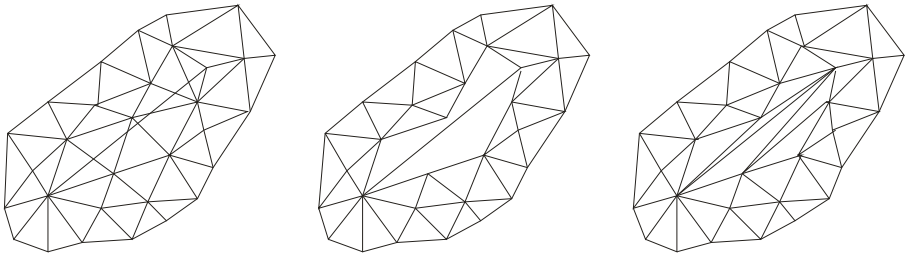
Рассматривавшаяся выше задача вывода изображения топографической поверхности сейчас может быть уточнена. Она еще сложнее, чем было показано. Генерализация изображения модели поверхности должна осуществляться с учетом генерализации изображения элементов ситуации.

Объекты ситуации, являющиеся одновременно и объектами топографической поверхности, представляют собой чаще всего линейные или полосные объекты, реже – площадные (терриконы, котлованы, водоемы…). Если, например, изображение площадного или полосного объекта в результате генерализации исчезает, то изображение горизонталей должно оставаться корректным, они не должны в таких местах прерываться.

В целом задача согласования представления топографической поверхности с представлением ситуации в процессе генерализации или без таковой сводится

все к той же задаче перестройки триангуляции или построению неявной триангуляции.

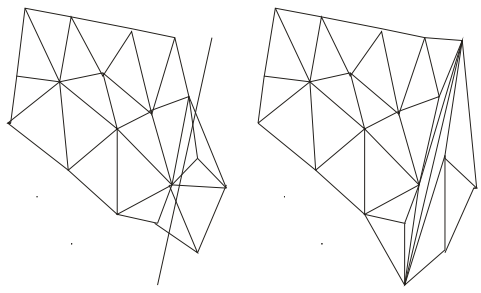
Пусть имеется некоторая триангуляция, на которую накладывается линейный объект (рис. 8.94). Предположим, что начальная и конечная точки этого объекта совпадают с двумя вершинами триангуляции. Если начало и конец отрезка не совпадают с вершинами триангуляции, то определяются треугольники, в которые они попадают, и триангуляция корректируется. Если начальная и конечная точка вставляемого линейного объекта имеют значения высоты, то они включаются в триангуляцию как новые вершины. Если для какой-либо вершины значение высоты неизвестно, то оно вычисляется по существующей триангуляции. Следовательно, можно считать, что начало и конец линейного объекта совпадают с двумя вершинами триангуляции.



а) *б*) *в*)

Рис. 8.94. Вставка линейного объекта

Тогда, чтобы включить линейный объект в триангуляцию, удаляются все пересекаемые объектом ребра и инцидентные им треугольники. В результате образуется внутренняя область, рассекаемая линейным объектом на левую и правую подобласти (рис. 8.94, *б*). Рассмотрим для определенности правую подобласть. Будем двигаться по ее границе так, чтобы подобласть оставалась слева, то есть будем совершать ее обход против часовой стрелки. За начало движения примем граничную вершину подобласти, смежную с началом линейного объекта. Если следующие друг за другом вершины v_{i-1}, v_i и v_{i+1} образуют треугольник, лежащий слева по ходу движения, то создаются ребро (v_{i-1}, v_{i+1}) и треугольник (v_{i-1}, v_i, v_{i+1}), после чего вершина v_i исключается из списка граничных вершин подобласти. Процесс перестроения триангуляции в правой подобласти заканчивается, когда в списке ее граничных вершин останется одна вершина. Эта вершина соединяется ребрами с начальной и конечной точками линейного объекта и создается последний треугольник. Таким образом, перестройка триангуляции в правой подобласти завершена.



а) *б*)

Рис. 8.95. Точки вне области

Корректировка триангуляции в левой подобласти выполняется совершенно таким же образом, если двигаться не от начальной точки отрезка, а от конечной точки к

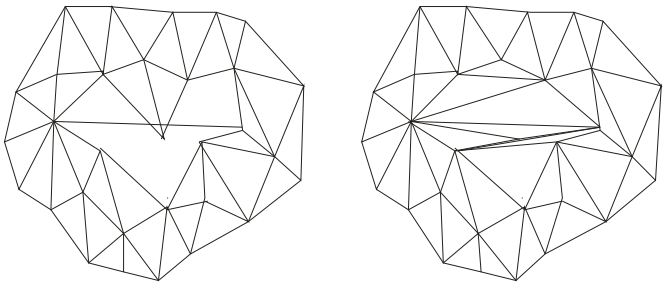
начальной. Тогда левая подобласть станет «правой» и все перечисленные правила сохраняются.

В реальных задачах могут встретиться случаи, когда одна из точек линейного объекта или обе точки находятся вне области триангуляции (рис. 8.95). Также возможно пересечение линейным объектом триангуляции с дырами. Программа для решения этой задачи должна обрабатывать все возможные случаи, начиная с того, когда обе точки попадают в один треугольник.

Таким же образом в триангуляцию могут включаться структурные линии. Первоначально триангуляция создается без их учета. После этого каждый отрезок структурной линии вставляется в созданную триангуляцию, и она перестраивается таким же образом, как было описано выше. Разумеется, точки структурных линий совпадают с вершинами триангуляции. Такой способ построения триангуляции в алгоритмическом отношении намного проще, чем учет структурных линий непосредственно в процессе ее создания (за исключением волновых алгоритмов построения триангуляции).

Триангуляцию, в которую включены структурные линии, ее границы либо другие объекты, называют *триангуляцией с ограничениями*. Ограничение состоит в том, что структурные и другие линии не должны пересекаться сторонами треугольников, а могут только совпадать с ними.

При наложении на триангуляцию двумерного объекта из нее исключаются все вершины, попадающие внутрь этого объекта, и все инцидентные им ребра (рис. 8.96). После этого для каждой стороны многоугольника, описывающего двумерный объект, осуществляется перестройка триангуляции в примыкающей снаружи подобласти так, как было описано для линейного объекта. Внутри этого объекта триангуляция, в зависимости от требований, может строиться или не строиться. Если такая триангуляция строится, то в нее включаются все граничные вершины объекта и его внутренние точки, если таковые имеются. (Примером может служить терриконик.)



а) *б*)

Рис. 8.96. Вставка в «дыру»

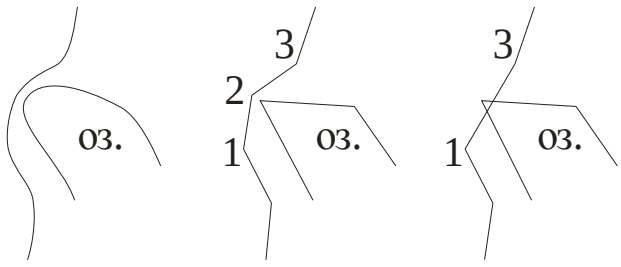


Рис. 8.97. Нарушение топологии

При разработке программ генерализации ситуации необходимо учитывать, что в процессе генерализации возможны нарушения топологии. На рис. 8.97 слева изображен

фрагмент ситуации, на котором линейный объект (дорога) огибает площадной (озеро). На том же рис. 8.97 в центре показаны ломаные линии, представляющие каждый из этих объектов. В процессе генерализации для вывода изображения в уменьшенном масштабе вершина 2 может быть исключена из списка вершин на том основании, что ее отклонение от прямой, проходящей через вершины 1 и 3, меньше величины δ для этого масштаба. Тогда мы получим картину (на рис. 8.97 справа), на которой можно видеть явную топологическую ошибку – пересечение озера дорогой.

Каждая подобная ситуация должна быть проанализирована и, если есть необходимость, нарушение топологии должно быть исправлено. Для этого достаточно включить вершины, подобные вершине 2, в список вершин, отображаемых в уменьшенном масштабе.

На рис. 8.98 представлен нестандартный участок дороги, проходящей по озеру. Такие участки существуют в действительности, и их длина может достигать сотен метров. Подобные объекты представляют своеобразную патологию, но как раз рассмотрение и моделирование патологических случаев позволяют больше узнать и об обычных объектах. Приведенный пример позволяет продемонстрировать, как разработчики программного обеспечения и пользователи понимают некоторые вопросы моделирования.

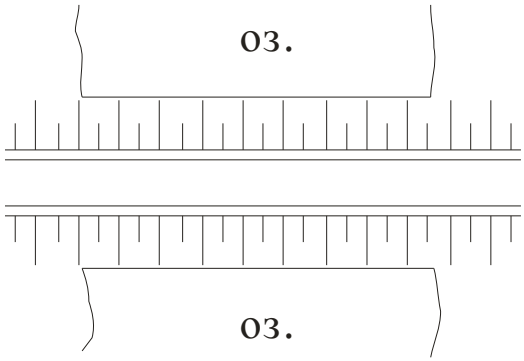
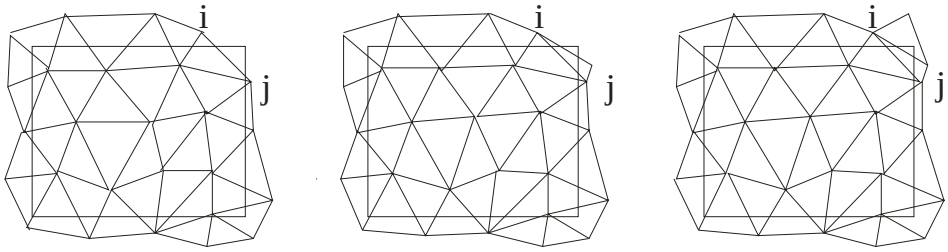


Рис. 8.98. Пример ситуации

При создании модели топографической поверхности в нее должны быть включены, как минимум, точки, расположенные по низу откоса дорожной насыпи. В идеале она должна содержать также и точки по верхней бровке насыпи. Однако, на практике чаще всего создаются «цифровые карты», когда стремятся получить нечто близкое к картографическому изображению. При таком понимании геоинформационной модели в нее включат дорогу вместе с насыпью и озера. В итоге при использовании модели можно будет получить удовлетворительное изображение для конкретного масштаба карты. Но трудно сказать, что будет при генерализации геоинформационной модели.

При выводе триангуляции на экран монитора в его углах возникают некоторые особенности. Чтобы вывести изображение триангуляции на экран, необходимо выбрать все вершины триангуляции, попадающие в область экрана, и все смежные с ними вершины, после чего отобразить все инцидентные им ребра и треугольники. Такое решение позволит почти всегда получить корректное изображение. Но в некоторых случаях будут возникать ошибки, подобные показанной на рис. 8.99, *а*. Часть треугольника, в который попадает правый верхний угол экрана, изображена не будет. Чтобы избавиться от этой ошибки, необходимо проверить, в какой треугольник попадает каждый угол экрана. Если все вершины треугольника находятся среди отобранных, то

изображение будет корректным. На рис. 8.99, *б* видно, что если отобразить второй (внешний) треугольник, инцидентный ребру (i, j), то этого может оказаться недостаточным. Таким образом, в подобных случаях может потребоваться добавление нескольких треугольников, чтобы изображение было корректным (рис. 8.99, *в*).



а) *б*) *в*)

Рис. 8.99. Дефект изображения

Содержание данного раздела свидетельствует о том, что решения задачи построения статической триангуляции иногда недостаточно. В общем случае задача создания модели топографической поверхности должна решаться совместно с задачей моделирования объектов топографической ситуации. Как мы надеемся, нам удалось показать необходимость согласования модели поверхности с моделью ситуации. Вообще же можно заметить, что построение моделей топографических поверхностей зависит от множества задач, для решения которых эти модели создаются. Так, если мы хотим решать инженерные задачи с применением модели топографической поверхности, то такая модель должна содержать все вершины, принадлежащие поверхности, и ничего больше. Но если нам необходимо, например, вычислять поправки за рельеф при создании ортофотопланов, то очевидно, что наша модель должна покрывать всю область моделирования, включая те участки земной поверхности, на которых топографическая поверхность «не существует»: здания, водоемы и т. п.

8.24. Моделирование неоднозначных поверхностей

Задача моделирования неоднозначных поверхностей возникает в связи с необходимостью представления поверхностей, не являющихся однозначными функциями двух переменных: обрывов, подпорных стенок, нависающих скал и т. п. Раньше данная задача возникала при обработке материалов фототеодолитной съемки в горных районах, теперь может возникать при обработке результатов съемки с помощью наземных систем лазерного сканирования. На рис. 8.100 представлен фрагмент триангуляции, уложенной на неоднозначной поверхности. Треугольники, расположенные на рис. 8.100 ниже ломаной (1, 2, 3, 4), принадлежат обратному склону. Справа на этом рисунке показан вертикальный профиль по линии *AB*.

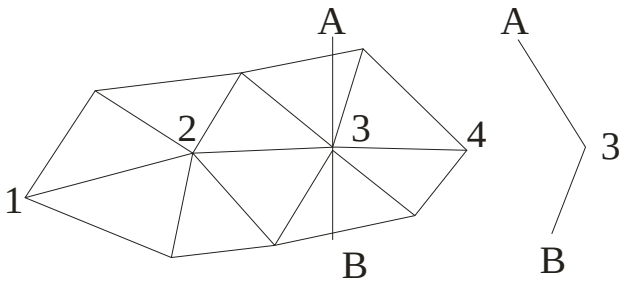


Рис. 8.100. Неоднозначная поверхность

Неоднозначные поверхности подтверждают преимущества использования моделей на нерегулярной сетке треугольников, поскольку *треугольники могут быть уложены на любой поверхности*. Уложить же на любой поверхности сетку четырехугольников, хотя бы и произвольных, нельзя, поскольку возможно вырождение некоторых из них в треугольники.

Для построения триангуляции на неоднозначной поверхности требуется использование понятий ориентации и ориентируемой поверхности. Исходным понятием при этом служит ориентация прямой. По прямой можно двигаться в одном из двух противоположных направлений. Если направление движения по прямой указано, то о ней говорят, как об *ориентированной прямой* или, что *прямая ориентирована*.

Аналогичным образом вводится понятие разомкнутой *ориентированной кривой*. Если кривая замкнута, то ее можно ориентировать либо по часовой стрелке, либо против часовой стрелки.

На основе понятия ориентированной замкнутой кривой вводится понятие ориентированной плоскости. Если некоторый участок плоскости ограничен простой замкнутой кривой, то, как только что было сказано, ее можно ориентировать двумя способами. Считается, что при ориентации такой кривой ориентируется и ограничиваемый ею участок плоскости. При движении по ориентированной замкнутой кривой ограниченный кусок плоскости остается все время либо слева (при обходе против часовой стрелки), либо справа (при движении по часовой стрелке). Чтобы задать на плоскости ориентацию всевозможных замкнутых кривых, достаточно указать ориентацию хотя бы одной замкнутой кривой; ориентация остальных замкнутых кривых на ней считается такой же. Плоскость вместе с заданной на ней ориентацией замкнутых кривых называется *ориентированной плоскостью*. Очевидно, что плоскость может быть ориентирована двумя способами.

Ориентация плоскости может быть указана также путем выбора системы координат. Выбирая систему координат, мы устанавливаем знак углов, расположенных на плоскости. При этом устанавливается и знак площадей, ограниченных замкнутыми кривыми. На рис. 8.101 слева показана левая система координат, справа – правая. При вычислении площади, ограниченной простой замкнутой кривой *C*, по формуле криволинейного интеграла

$$S = \frac{1}{2}\int_C x dy - y dx$$

в указанном стрелкой направлении будет получено положительное значение, при

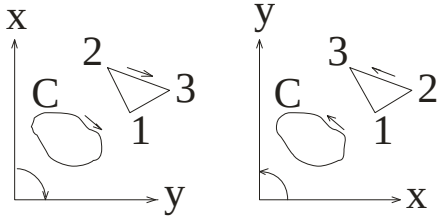


Рис. 8.101. Ориентация плоскости

вычислении площади в обратном направлении – отрицательное.

Таким же образом удвоенная площадь треугольника, вычисляемая по формуле

$$2S = \sum_{i=1}^{3} x_i (y_{i+1} - y_{i-1}) \qquad (8.235)$$

в направлении, указанном стрелкой, будет положительна, а вычисляемая в обратном направлении – отрицательна.

Любая поверхность, ограничивающая часть пространства (сфера, эллипсоид, многогранник) также может быть ориентирована. Ориентация поверхности устанавливается указанием направления обхода кусков поверхности, ограниченных замкнутыми кривыми. Два ограниченных участка поверхности называют *ориентированными одинаково*, если при их обходе по ограничивающим кривым в заданном направлении участки остаются с одной и той же стороны. Поверхность, для которой определена ориентация ее кусков, ограниченных замкнутыми кривыми, называется *ориентированной поверхностью*.

Наряду с ориентируемыми поверхностями существуют и *неориентируемые*. Наиболее известным примером такой поверхности является лист Мебиуса (рис. 8.102), который, кроме того, является еще и *односторонней поверхностью*.

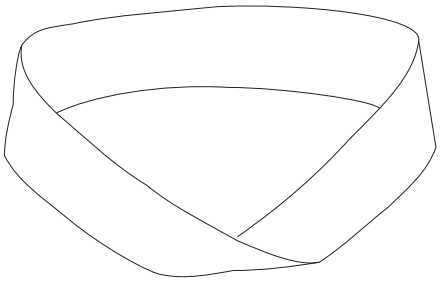


Рис. 8.102. Лист Мебиуса

Теперь мы можем вернуться к представлению неоднозначных поверхностей. Треугольники, уложенные на неоднозначной поверхности, должны быть одинаково ориентированными, желательно в положительном направлении. Ориентация треугольников означает упорядочивание их вершин таким образом, что при обходе вершин в порядке их перечисления треугольники всегда остаются с одной стороны. Другим способом указания ориентации треугольников является приписывание каждому треугольнику некоторого признака ориентации, который может принимать всего два значения. Но такой способ хуже, поскольку требует дополнительной памяти для хранения значения признака ориентации.

Если моделируемая поверхность является однозначной, то для определения ориентации треугольника достаточно вычислить площадь его проекции на поверхность относимости. Если полученное значение площади отрицательно, то изменить ориентацию треугольника можно, поменяв местами его первую и третью вершины.

Если моделируемая поверхность является заведомо неоднозначной или мы допускаем такую возможность, то для ориентации всех треугольников достаточно задать явным образом ориентацию *одного* треугольника.

Ориентация остальных треугольников может быть определена программным путем.

На рис. 8.103 представлен фрагмент ориентированной триангуляции. Нетрудно видеть, что если все треугольники ориентированы одинаково, то обход каждого ребра в двух смежных треугольниках осуществляется в противоположных направлениях. Отсюда следует, что, имея ориентацию хотя бы одного треугольника, ориентацию остальных можно установить без вычисления значения их площади.

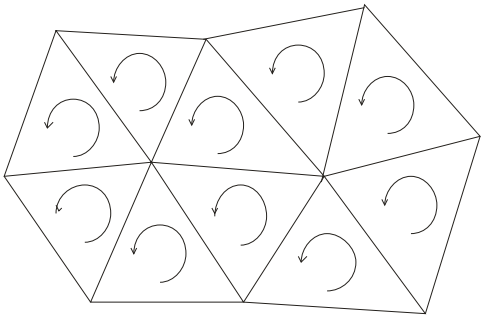


Рис. 8.103. Ориентация треугольников

Рассмотрим теперь, что будет происходить с треугольниками на неоднозначной поверхности. На рис. 8.104 треугольник (1, 2, 3) находится на обычном склоне, а треугольники (1, 3, 4) и (1, 4, 5) – на обратном. Проекции вершин триангуляции имеют те же номера со штрихом. Первое, что можно заметить на рис. 8.104, это то, что проекции треугольников перекрываются. Следовательно, проекция триангуляции на неоднозначной поверхности не является разбиением области моделирования и может не быть плоской триангуляцией.

Кроме того, проекции треугольников на обратных склонах меняют свою ориентацию. Отсюда следует, что если все треугольники имеют положительную ориентацию, то площадь проекции треугольника на обратном склоне, вычисляемая по формуле (8.235), будет отрицательной. Данное обстоятельство может использоваться для обнаружения в уже построенной триангуляции треугольников на обратных склонах программным путем.

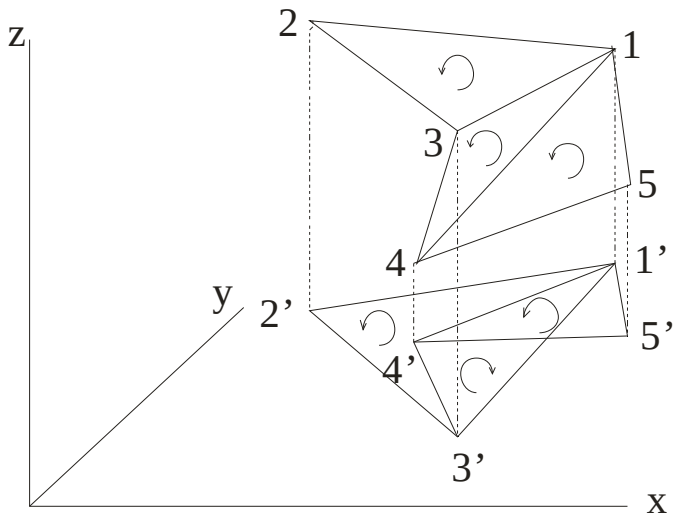


Рис. 8.104. Проекция треугольников

Также очевидно, что некоторые особенности возникают при моделировании подпорных стенок или искусственных котлованов с вертикальными стенками. Если их проекции на горизонтальную поверхность образуют прямые линии, то никаких проблем с их моделированием не возникает. Необходимо только, чтобы в нужных местах вершины триангуляции располагались на одних и тех

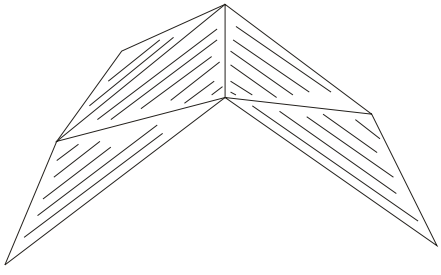


Рис. 8.105. Подпорная стенка

же вертикальных линиях. На рис. 8.105 представлен пример подпорной стенки. Треугольники лежат в двух вертикальных плоскостях, образующих подпорную стенку. Площадь проекции таких треугольников, вычисленная по формуле (8.235), будет иметь нулевое значение. Следовательно, разбиение вертикальной плоскости на треугольники может осуществляться разными способами, что подтверждает и рис. 8.105.

Если соблюдать указанное выше требование о расположении точек на вертикальных линиях, то при моделировании вертикальных поверхностей более сложной формы никаких дополнительных проблем не возникает.

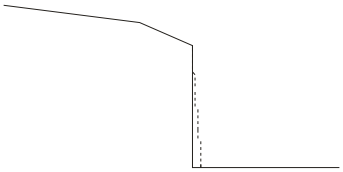


Рис. 8.106. Исключение неоднозначности

При построении моделей вертикальных поверхностей можно также использовать такой прием, как изменение координат на незначительную величину, например, на 0.01 м, что позволяет избавиться от неоднозначности функции двух переменных. Данный прием в утрированном виде показан на вертикальном профиле на рис. 8.106.

8.25. Методы сглаживания

Методы сглаживания применяются, когда известно, что значения высоты или глубины в исходных точках получены с существенными ошибками. В некоторых случаях методы сглаживания имеют преимущества перед интерполированием. Однако, при моделировании топографических поверхностей эти методы нуждаются в критическом отношении. При топографических съемках, как правило, стремятся минимизировать объем исходных данных при условии соблюдения заданной точности. Это особенно справедливо для случаев, когда снимаются характерные точки и структурные линии топографической поверхности. Однако сглаживающие методы работают таким образом, что большое отклонение точки от некоторого среднего положения воспринимается ими как свидетельство наличия существенной ошибки в данной точке. В результате высоты точек искажаются тем сильнее, чем характернее они являются.

Аппроксимация может применяться и в тех случаях, когда высоты (глубины) измерены точно. Но тогда необходимо принимать меры по ослаблению отмеченного выше эффекта искажения характерных точек. Это можно сделать, если значениям высот придавать тем большие веса, чем больше отклонение точки. Хотя ухищрения такого рода могут несколько улучшить представление поверхности, они не заслуживают массового применения.

Фактически при топографических съемках измерения выполняются с очень высокой точностью. Уровень ошибки в самых худших случаях на один порядок меньше колебаний высот; в среднем он меньше на 2–3 порядка. Применение сглаживающих методов приводит к нивелированию поверхности, что вступает в конфликт с разумной традицией подчеркивать выдающиеся элементы топографической поверхности при составлении топографических карт и планов. По этой причине сглаживающие методы при моделировании

топографических поверхностей получили меньшее распространение, чем методы интерполирования.

Необходимыми предпосылками применения методов сглаживания служат:

– случайный характер точек, представляющих поверхность;

– достаточно большое значение отношения ошибки измерения высот к диапазону изменения высот;

– высокая, и даже избыточная, плотность исходных точек.

Все три фактора обычно имеют место при съемке дна водоемов и морских акваторий, когда судно ходит галсами. Измерение глубин по каждому галсу осуществляется с высокой частотой (малым шагом между соседними точками), поскольку осуществляется в автоматическом режиме с помощью эхолота.

Наиболее популярными принципами сглаживания являются:

– минимизация наибольшего абсолютного отклонения аппроксимирующей функции от исходных точек

$$\max \left| H(P_i) - z_i \right| = \min , \qquad (8.236)$$

называемая *Чебышевской аппроксимацией*;

– минимизация суммы

$$\sum_{i=1}^{n} (H(P_i) - z_i)^2 = \min, \qquad (8.237)$$

называемая *аппроксимацией по методу наименьших квадратов*.

В формулах (8.236) и (8.237) n – число исходных точек; z_i – высота топографической поверхности в точке $P_i = (x_i, y_i)$; $H(P_i)$ – значение аппроксимирующей функции в точке P_i . Использование Чебышевской аппроксимации позволяет представить поверхность так, что ошибка не превосходит определенного предела. Сглаживание по методу наименьших квадратов не гарантирует, что в какой-либо точке ошибка аппроксимации не превысит установленного допуска. Данный критерий сглаживания характеризует приближение к поверхности в целом. Другие оценки точности аппроксимирующей функции можно найти в работе [9].

Метод наименьших квадратов чаще всего служит основой для конструирования регулярной дискретной модели способом динамической поверхности. Данный способ назван так потому, что с изменением координат x и y изменяются параметры уравнения, описывающего поверхность. Если проводить сравнение различных методов моделирования топографической поверхности, то надо отметить следующее:

– при создании непрерывных моделей аналитическое выражение описывает всю топографическую поверхность в целом;

– при применении кусочно-непрерывных моделей аналитическое выражение описывает лишь отдельный элемент топографической поверхности;

– при использовании способа динамической поверхности аналитическое выражение характеризует топографическую поверхность лишь в отдельно взятой точке.

Поскольку метод наименьших квадратов часто используется при построении динамической поверхности, приведем его описание на примере сглаживания функций одной переменной.

Пусть некоторая функция задана своими значениями y_i в узлах x_i $(i=1,...,n)$, причем известно, что значения функции содержат ошибки ε_i. Из тех или иных соображений может быть также известно, что функция имеет общий вид $F(x)$, но неизвестны значения параметров выражения, описывающего данную конкретную функцию. Поэтому мы можем рассматривать $F(x)$ как функцию от неизвестных $m < n$ параметров

$$y = F(x, a_1, ..., a_m),$$

которые требуется определить по исходным данным поставленной задачи.

В соответствии с методом наименьших квадратов наилучшими значениями параметров a_i считаются те, что удовлетворяют условию

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \min$$

или

$$S = \sum_{i=1}^{n} (F(x_i, a_1, ..., a_m) - y_i)^2 = \min, \qquad (8.238)$$

то есть сумма квадратов отклонений полученной функции от значений функции в исходных точках минимальна.

Целевая функция $S(a_1, ..., a_m)$ является функцией многих переменных. Необходимым условием экстремума функции многих переменных является равенство нулю ее частных производных:

$$\frac{\partial S}{\partial a_j} = 0, \ (j=1,...,m).$$

Тогда, продифференцировав выражение (8.238) по неизвестным a_j, получим *m* уравнений вида

$$\sum_{i=1}^{n} (F(x_i, a_1, ..., a_m) - y_i) \frac{\partial F(x_i, a_1, ..., a_m)}{\partial a_j} = 0 \ \ (j=1,...,m). \quad (8.239)$$

Полученную систему уравнений называют *системой нормальных уравнений*. Нормальные уравнения имеют наиболее простой вид тогда, когда функция $y = F(x, a_1, ..., a_m)$ линейна относительно неизвестных параметров a_j.

Напомним, что при моделировании топографических поверхностей наиболее часто приближение функций одной или двух переменных представляется как линейная комбинация функций. В частности, в случае функций одной переменной, который мы рассматриваем только для определенности, линейная комбинация функций представляется выражением

$$F(x, a_1, ..., a_m) = a_1 f_1(x) + ... + a_m f_m(x).$$

Следовательно, уравнения (8.239) можно конкретизировать и записать как

$$\frac{\partial S}{\partial a_j} = \sum_{i=1}^{n} [a_1 f_1(x_i) + ... + a_m f_m(x_i) - y_i] f_j(x_i) = 0 \ \ (j = 1, ..., m), \quad (8.240)$$

или в развернутом виде

$$\left.\begin{aligned} &\frac{\partial S}{\partial a_1} = \sum_{i=1}^{n} [a_1 f_1(x_i) + ... + a_m f_m(x_i) - y_i] f_1(x_i) = 0 \\ &\frac{\partial S}{\partial a_2} = \sum_{i=1}^{n} [a_1 f_1(x_i) + ... + a_m f_m(x_i) - y_i] f_2(x_i) = 0 \\ &\qquad\qquad ... \\ &\frac{\partial S}{\partial a_m} = \sum_{i=1}^{n} [a_1 f_1(x_i) + ... + a_m f_m(x_i) - y_i] f_m(x_i) = 0 \end{aligned}\right\}. \quad (8.241)$$

Значения y_i перенесем в правую часть и получим

$$\left.\begin{aligned} &\frac{\partial S}{\partial a_1} = \sum_{i=1}^{n} [a_1 f_1(x_i) + ... + a_m f_m(x_i)] f_1(x_i) = \sum_{i=1}^{n} f_1(x_i) y_i \\ &\frac{\partial S}{\partial a_2} = \sum_{i=1}^{n} [a_1 f_1(x_i) + ... + a_m f_m(x_i)] f_2(x_i) = \sum_{i=1}^{n} f_2(x_i) y_i \\ &\qquad\qquad ... \\ &\frac{\partial S}{\partial a_m} = \sum_{i=1}^{n} [a_1 f_1(x_i) + ... + a_m f_m(x_i)] f_m(x_i) = \sum_{i=1}^{n} f_m(x_i) y_i \end{aligned}\right\}.$$

Перепишем полученную систему уравнений, сгруппировав произведения, содержащие $f_j(x_i)$:

$$\left.\begin{aligned} &\sum_{i=1}^{n} a_1 f_1(x_i) f_1(x_i) + ... + \sum_{i=1}^{n} a_m f_m(x_i) f_1(x_i) = \sum_{i=1}^{n} f_1(x_i) y_i \\ &\sum_{i=1}^{n} a_1 f_1(x_i) f_2(x_i) + ... + \sum_{i=1}^{n} a_m f_m(x_i) f_2(x_i) = \sum_{i=1}^{n} f_2(x_i) y_i \\ &\qquad\qquad ... \\ &\sum_{i=1}^{n} a_1 f_1(x_i) f_m(x_i) + ... + \sum_{i=1}^{n} a_m f_m(x_i) f_m(x_i) = \sum_{i=1}^{n} f_m(x_i) y_i \end{aligned}\right\}.$$

Неизвестные коэффициенты можно вынести за знак суммы, после чего придем к выражениям

$$\left.\begin{aligned}
&a_1\sum_{i=1}^{n} f_1(x_i)\,f_1(x_i)+\ldots+a_m\sum_{i=1}^{n} f_m(x_i)\,f_1(x_i)=\sum_{i=1}^{n} f_1(x_i)\,y_i\\
&a_1\sum_{i=1}^{n} f_1(x_i)\,f_2(x_i)\,y_i+\ldots+a_m\sum_{i=1}^{n} f_m(x_i)\,f_2(x_i)=\sum_{i=1}^{n} f_2(x_i)\,y_i\\
&\qquad\qquad\ldots\\
&a_1\sum_{i=1}^{n} f_1(x_i)\,f_m(x_i)\,y_i+\ldots+a_m\sum_{i=1}^{n} f_m(x_i)\,f_m(x_i)=\sum_{i=1}^{n} f_m(x_i)\,y_i
\end{aligned}\right\}.$$

Введем обозначения для сумм

$$s_{jk}=\sum_{i=1}^{n} f_j(x_i)\,f_k(x_i)\ \ (j=1,\ldots,m)\ \ (k=1,\ldots,m); \qquad (8.242)$$

$$t_j=\sum_{i=1}^{n} f_k(x_i)\,y_i\quad (k=1,\ldots,m). \qquad (8.243)$$

Тогда систему нормальных уравнений можно представить в окончательном виде

$$\left.\begin{aligned}
&s_{11}a_1+\ldots+s_{1m}a_m=t_1\\
&s_{21}a_1+\ldots+s_{2m}a_m=t_2\\
&\qquad\ldots\\
&s_{m1}a_1+\ldots+s_{mm}a_m=t_m
\end{aligned}\right\}. \qquad (8.244)$$

Матрица коэффициентов системы нормальных уравнений является плотно заполненной, поэтому для ее решения требуются значительные вычислительные затраты. Кроме того, из приведенных формул видно, что затраты необходимы для получения самой матрицы коэффициентов. Громоздкость вычислений считают одним из недостатков метода наименьших квадратов. Но он обладает тем свойством, что если сумма S квадратов отклонений ε мала, то и значения ε также будут малы по абсолютной величине.

Определитель системы (8.244) называют *определителем Грамма* и при некоторых условиях он может быть малой величиной, поэтому вычисления могут стать неустойчивыми относительно погрешностей в исходных данных и погрешностей округления.

Сущность *способа динамической поверхности* состоит в следующем. Пусть задано множество исходных точек $\{P_i : i=1,\ldots,|P|\}$. Для определения значения аппроксимирующей функции в произвольной точке (x, y) из всего множества исходных точек отбираются только точки P_i, отвечающие условию

$$((x-x_i)^2+(y-y_i)^2)\le R^2,$$

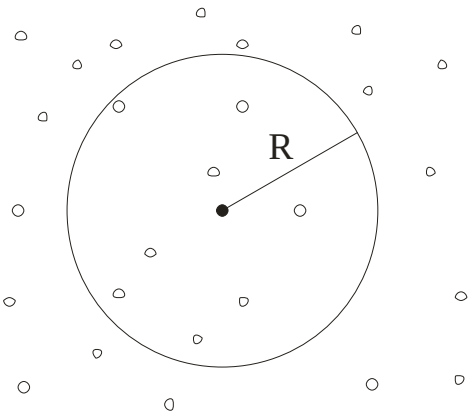


Рис. 8.107. Динамическая поверхность

где R – некоторый фиксированный радиус (рис. 8.107).

Далее предполагается, что в определяемой точке топографическая поверхность описывается уравнением $H(x,y)$. Обычно в качестве такого уравнения выбирается

– уравнение горизонтальной плоскости

$$H(x,y)=c;$$

– уравнение наклонной плоскости

$$H(x,y)=ax+by+c;$$

– уравнение поверхности второго порядка, например,

$$H(x,y)=ax^2+bxy+cy^2+dx+ey+f.$$

Положение динамической поверхности определяется с помощью обычной техники метода наименьших квадратов:

– устанавливается критерий

$$\sum_{i=1}^{m} p_i\left(H(x_i,y_i)-z_i\right)^2=\min,$$

где $m\le n$ – число исходных точек, удаленных от определяемой не более чем на фиксированное расстояние R; p_i – вес точки P_i – является некоторой неотрицательной убывающей функцией от R: $p_i=\frac{1}{R_i}$, $p_i=\frac{1}{R_i+c}$, $p_i=\frac{1}{R_i^2+c}$ и т. п., где $c>0$ – некоторая константа;

– составляется и решается система нормальных уравнений, в которой неизвестными являются коэффициенты уравнения.

Конкретные реализации метода динамической поверхности отличаются друг от друга выбором радиуса R, видом весовых функций p_i и видом уравнения аппроксимирующей поверхности [32]. Надо признать, что при всей его математической «нестрогости», метод динамической поверхности покоится на вполне здравых основаниях. Поскольку топографическая поверхность обладает плохими дифференциальными свойствами, то вполне естественно предполагать, что влияние исходных точек убывает по мере их удаления от определяемой точки, и исходные точки, достаточно удаленные от определяемой, никаким образом не оказывают влияния на ее высоту и могут игнорироваться. Поэтому для вычисления высоты в некоторой точке вполне достаточно привлечения ближайших к ней исходных точек.

Но применение способа динамической поверхности не является исключением в том смысле, что также связано с некоторыми затруднениями. В частности, при его использовании желательно сравнительно равномерное распределение исходных точек по всей области моделирования. При значительном изменении плотности исходных точек в пределах одного моделируемого участка могут возникнуть трудности с определением радиуса R.

Малое значение R может привести к тому, что для некоторой определяемой точки (x, y) в круг заданного радиуса не попадет ни одна исходная точка либо попадет число точек, недостаточное для определения параметров уравнения аппроксимирующей поверхности. Иногда этот недостаток устраняется либо понижением степени сглаживающего полинома, либо увеличением радиуса R (только для данной точки).

Слишком большое значение R может привести:

– во-первых, к заметному возрастанию затрат машинного времени;

– во-вторых, к некоторому снижению точности моделирования, поскольку в общем случае далеко расположенные точки искажают значение высоты в определяемой точке.

Для некоторых определяемых точек при использовании способа динамической поверхности может случиться так, что все отобранные исходные точки окажутся не более или менее равномерно распределенными по горизонту, а лежащими по одну его сторону. На рис. 8.108 такие точки отмечены знаком «+». В подобных случаях оказывается, что решается *задача экстраполяции*, тогда как мы предполагали решение задачи интерполяции. Примером проблемной области, где постоянно приходится решать задачу экстраполяции, является прогнозирование погоды, когда по значениям температуры, давления, влажности, скорости и направления ветра за предыдущий период требуется определить их значения в последующие дни, недели и т. д. Качество прогнозов хорошо известно. И точность прогнозов, как правило, резко падает по мере возрастания срока прогнозирования. Это при том, что, по статистике, прогноз в двух случаях из трех будет правильным, если на завтра предсказывать сегодняшнюю погоду.

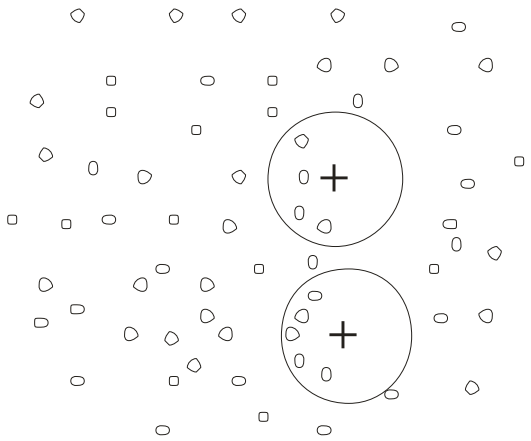


Рис. 8.108. Точки экстраполяции

Задача экстраполяции является более сложной математической задачей, чем интерполяция. И ее более или менее корректное решение возможно лишь для функций с хорошими дифференциальными свойствами, для функций с «предсказуемым» поведением. Но естественный рельеф земной поверхности к таким функциям не относится, поэтому его экстраполяция может сопровождаться появлением грубых ошибок в созданной информационной модели.

Метод динамической поверхности хорош тем, что он позволяет решать некоторые задачи на топографической поверхности без создания ее информационной модели. Точнее сказать, информационной моделью топографической поверхности может служить само множество исходных точек, расположенных в характерных точках этой поверхности и представляющих

собой неструктурированную или плохо организованную модель. В таких случаях может быть использована *виртуальная* регулярная дискретная модель топографической поверхности. Такая модель может представлять собой сетку квадратов, но в памяти ЭВМ хранятся только ее параметры $\{X_0, Y_0, Z_0, S_x, S_y, S_z, L_x, L_y, \alpha\}$ (координаты угла, начальная высота, величина шага и число узлов по каждой координате, угол разворота относительно системы координат), а значения высот ее узлов в памяти не хранятся и вычисляются по мере необходимости.

Однако неупорядоченное хранение большого числа исходных точек потребует много времени на вычисления, так как для каждого узла виртуальной дискретной модели будет требоваться просмотр всех исходных точек с целью отбора попадающих в круг радиуса *R*. Поэтому использование виртуальных дискретных моделей требует упорядочивания исходных точек с применением квадродеревьев либо других способов. Виртуальная модель может создаваться для решения некоторых редких задач. Если требуется решать те или иные задачи часто, то целесообразнее создать модель один раз и хранить ее во внешней памяти.

Еще одним недостатком виртуальной модели является то, что ее нельзя исправить, поскольку она не хранится (по определению) не только во внешней, но и во внутренней памяти.

Построение сглаживающих поверхностей выполняется с применением всех трех известных методов представления топографических поверхностей: непрерывной, кусочно-непрерывной и динамической поверхности [6].

По сравнению с методом интерполяции, сглаживание непрерывной поверхностью дает возможность уменьшить число неизвестных и сократить вычислительные затраты. Но применение сглаживания сопровождается определенным риском; необходимые условия его применения перечислялись выше. Тем не менее, полностью избежать недостатков, присущих интерполированию с помощью обобщенных полиномов, при переходе к представлению топографической поверхности сглаживающей непрерывной функцией не удается.

Представление топографической поверхности сглаживающей кусочно-непрерывной функцией двух переменных наряду со снижением требований к объему памяти ЭВМ и сокращением времени вычислений позволяет получать более точные модели, чем непрерывные.

К данной группе методов относится *r-гладкое приближение функций*, использовавшееся для представления геофизических полей и описанное в работе [5]. Область определения функции покрывается множеством перекрывающихся квадратов с фиксированной стороной 2*s*, центры которых совпадают с нерегулярно расположенными исходными точками. Каждой исходной точке ставится в соответствие функция вида

$$\varphi_j(x,y)=\frac{\omega\left(\frac{x-x_j}{s},\frac{y-y_j}{s}\right)}{\sum_{i=1}^{n}\omega\left(\frac{x-x_i}{s},\frac{y-y_i}{s}\right)}, \qquad (8.245)$$

где

$$\omega=\begin{cases}\dfrac{(1-x^2)^{r+1}(1-y^2)^{r+1}}{(1-\lambda x^2)(1-\lambda y^2)} & -\text{если } |x|<1 \text{ и } |y|<1;\\ 0 & -\text{если } |x|\geq 1 \text{ или } |y|\geq 1.\end{cases} \qquad (8.246)$$

Первое выражение для ω используется, если $|x|<1$ и $|y|<1$, в противном случае $\omega=0$. Поверхность представляется выражением

$$H(x,y)=\sum_{i=1}^{n} z_i\varphi_i(x,y). \qquad (8.247)$$

Однако, такое решение может сопровождаться невязками в исходных точках:

$|H(x_i,y_i)-z_i|>\varepsilon$,

где ε – предельно допустимое значение *невязки* – отклонения вычисленной высоты от ее значения в исходной точке. Поэтому процесс вычисления значений функции в узлах сетки повторяется, пока максимальное отклонение не станет меньше допуска. В каждой итерации значения z_i заменяются полученными невязками. Качество аппроксимации зависит от выбора пользователем параметров *s*, *r*, λ.

Хотя такие попытки не предпринимались, но сглаживание топографических поверхностей могло бы выполняться с использованием методов, основанных на теории обобщенных сплайнов. Общее определение было дано в [21] Г.И. Марчуком: *сглаживающим сплайном* называют элемент $\sigma_\alpha\in x$, минимизирующий функционал вида

$$\Phi_\alpha(u)=\alpha\|Tu\|_4^2+\sum_{i=1}^{n}\left[(k_i,u)-z_i\right]^2,\ \alpha>0. \qquad (8.248)$$

Здесь $T:X\to Y$ – линейный оператор, действующий из X в Y (X, Y – гильбертовы пространства); k_i – линейно независимая система линейных ограниченных функционалов над пространством X; (k_i,u) – скалярное произведение; z_i – значения аппроксимируемой функции в опорных точках. Построение сглаживающего сплайна – это задача на отыскание абсолютного минимума и решается она иногда проще, чем поиск условного минимума, соответствующего задаче интерполирования. С практической точки зрения такой подход не может вызывать возражений, если отклонение можно сделать настолько малым, что им можно пренебречь. Кроме того, доказано, что при $\alpha\to 0$ сглаживающий сплайн сходится к интерполяционному.

Известен общий алгоритм построения сглаживающих и интерполяционных сплайнов [7], но его использование для аппроксимации функций двух переменных, заданных на нерегулярной сетке, сопряжено с серьезными трудностями. Выходом из положения является аппроксимация с помощью сплайнов на подпространстве, которая будет рассмотрена далее.

8.26. Сравнение способов конструирования поверхности

Конструирование *непрерывной модели* топографической поверхности обладает следующими свойствами.

1. В процессе получения уравнения непрерывной поверхности возникает необходимость решения больших систем линейных уравнений с плотно заполненными матрицами, что является серьезным недостатком.
2. Обусловленность систем уравнений ухудшается с возрастанием числа неизвестных, что требует специальных методов их решения и сопровождается увеличением времени вычислений.
3. Методы данной группы требуют сравнительно равномерного распределения исходных точек по области определения и весьма чувствительны к большим значениям первых производных.
4. Методы неприемлемы для представления негладких поверхностей.
5. В процессе получения непрерывной модели потребности в оперативной памяти могут быть очень большими. При хранении созданной модели требования к памяти минимальны, в лучшем случае необходимо хранить только значения коэффициентов уравнения, в худшем случае требуется сохранять еще и координаты исходных точек.
6. Точность аналитического выражения не всегда может быть удовлетворительной. Полученная поверхность может проходить через заданные точки, но между ними возможно появление осцилляций. Повышение точности путем увеличения плотности исходных точек и увеличения числа членов в выражении в некоторых случаях сопровождается обратными эффектами.
7. Важное преимущество непрерывных моделей – независимость от конфигурации моделируемой области и схемы выборки исходных точек. При регулярном распределении исходных точек удается повысить вычислительную эффективность методов, учитывая при определении коэффициентов факт периодичности распределения исходных точек. В некоторых случаях могут быть получены явные формулы.
8. При моделировании областей, не являющихся односвязными, возможно появление нежелательных осцилляций. Методы работают таким образом, что не делают различия между «дырами» в области определения и участками с разреженным распределением точек, поэтому каждая связная область должна моделироваться отдельно.
9. Вблизи границ области моделирования возможно появление заметных краевых эффектов.
10. Положительной стороной данных методов является их алгоритмичность. Затраты на реализацию того или иного метода сводятся к

разработке программы составления системы линейных уравнений, решение которой может осуществляться с помощью уже имеющихся программ.

11. Методы данной группы исключают возможность управления поведением моделирующей функции в некоторой локальной области.

12. Связывание в едином уравнении всех исходных точек по мере увеличения размеров моделируемой области или плотности точек теряет смысл, так как корреляционные зависимости между точками при этом ослабевают или исчезают вообще.

13. Описание границ области моделирования выглядит инородным элементом и требует дополнительной памяти.

14. Методы построения непрерывных поверхностей не применимы для представления неоднозначных поверхностей.

Построение модели топографической поверхности как кусочно-непрерывной выполняется более эффективно.

1. Отсутствует необходимость решения систем уравнений очень высокого порядка, вследствие чего снижаются требования к объему оперативной памяти, и уменьшается время вычислений; устраняются проблемы, связанные с плохой обусловленностью матриц.

2. Использование методов данной группы дает возможность значительно ослабить или полностью устранить нежелательные осцилляции, заметно повысить точность моделирования, в том числе, за счет разбиения топографической поверхности на более мелкие элементы.

3. Применение кусочно-непрерывных функций позволяет адекватно моделировать разрывные (не являющиеся гладкими) поверхности.

4. Плотность исходных точек может изменяться в широких пределах и устанавливаться индивидуально для каждого локального участка поверхности в зависимости от его сложности.

5. При использовании нерегулярной треугольной сетки легко преодолеваются неудобства, связанные с представлением моделируемых участков со сложными границами. Как правило, границы области моделирования совпадают с границами тех или иных треугольных элементов; исключение обычно составляют регулярные элементы.

6. Представление топографических поверхностей кусочно-непрерывными функциями дает возможность гибкого управления разбиением области моделирования на элементы. Их конфигурация и размеры могут либо жестко связываться с положением исходных точек, либо никак от них не зависеть.

7. Разбиение области моделирования на элементы позволяет ослабить краевые эффекты, характерные для любого способа конструирования поверхности.

8. Путем разбиения области моделирования на простейшие геометрические фигуры может быть снижена логическая сложность методов, имеющая место при разбиении области моделирования на произвольные элементы и одновременном повышении гладкости конструируемой поверхности. При использовании треугольных и четырехугольных сеток обеспечивается хорошая алгоритмичность методов моделирования.

9. Вследствие применения наиболее простых элементов оказывается возможным манипулировать положением поверхности в локальных областях.

10. Модели на нерегулярной сетке треугольников являются единственно возможными при моделировании неоднозначных поверхностей.

Методы построения непрерывных или кусочно-непрерывных поверхностей в виде крупных блоков неустойчивы к некоторым ошибкам в исходных данных. Если плановые координаты двух точек совпадают, то система уравнений либо противоречива (при разных значениях высот), либо ее определитель равен нулю (когда высоты имеют одинаковое значение).

Построение модели топографической поверхности *методом динамической* поверхности характеризуется следующими свойствами.

1. Требования к объему оперативной памяти минимальны. Данный метод может применяться для пересчета нерегулярной модели в регулярную. В таких случаях в оперативной памяти достаточно хранить множество исходных точек и строку или столбец регулярной модели.

2. Представление поверхностей с резкими формами или разрывных поверхностей связано с проблемами. В таких случаях в исходных данных необходимо задавать каким-либо образом линии разрыва гладкости или границы гладких областей.

3. Применение метода динамической поверхности возможно при сравнительно постоянной плотности исходных точек.

4. При реализации метода динамической поверхности необходимо предусматривать исключение возможности экстраполяции.

Сравнивая модели топографической поверхности разных типов между собой, можно отметить следующее:

– непрерывные модели не имеют практического значения;

– кусочно-непрерывные модели на нерегулярной сетке треугольников являются наиболее универсальным способом представления топографических поверхностей;

– регулярные кусочно-непрерывные модели являются наилучшим средством для решения некоторых инженерных задач, но мало пригодны для картографических целей.

8.27. Отображение дискретного множества на дискретное

Методы данного класса возникли в математике недавно и не отличаются таким разнообразием, как рассмотренные выше. Необходимо отметить, что здесь имеются в виду только прямые методы. Композиция методов – отображение дискретного множества на непрерывное, а затем отображение последнего на другое дискретное – к данному классу не относится. По этой причине, например, метод *r*-гладкого приближения функций и метод динамической поверхности были отнесены к отображениям предыдущего класса, поскольку в них так или иначе в явном виде присутствует аналитическое представление поверхности.

Используемые в настоящее время способы отображения дискретного множества на дискретное являются частными случаями метода конечных

элементов и появились они еще до того, как было получено аналитическое решение сплайн-интерполяции на множестве точек, произвольным образом расположенных в двумерной области. Основная идея этих способов заключена в том, что, если точное решение не может быть получено, то можно попытаться получить хорошее приближение к нему. С этой целью бесконечномерное функциональное пространство $\omega_2^m(\Omega)$ с достаточной степенью точности представляется конечномерным пространством E_h конечных элементов, и минимум функционала отыскивается на этом подпространстве. Решение поставленной задачи называют *сплайном на подпространстве*. Доказано, что сплайн на подпространстве при $h \to 0$ сходится к точному решению и, следовательно, аппроксимирует поверхность не хуже, чем точное решение сплайн-аппроксимации (разумеется, при надлежащем выборе E_h). Проблема заключается в выборе размеров элементов: при малых размерах элементов точность решения увеличивается, но одновременно увеличиваются и вычислительные затраты.

Особенностью метода является то, что область определения функции разбивается на конечные элементы прямоугольной или квадратной сеткой, никаким образом *не связанной* с исходными точками. В каждом узле сетки считается заданной *локальная функция* $\omega(x,y)$, равная 1 в этом узле и 0 – в остальных узлах. Функции $\omega(x,y)$ образуют *базис в пространстве конечных элементов* $\omega_2^m(\Omega)$. Далее решается задача либо интерполяции, либо сглаживания. Если при интерполяции отыскивается минимум функционала

$$\Phi(u) = \int_{\Omega} (u_x^2 + u_y^2)\,dxdy\,, \qquad (8.249)$$

то есть используются *билинейные конечные элементы*, то получают систему линейных уравнений вида

$$\begin{pmatrix} T & A \\ A & 0 \end{pmatrix} \begin{pmatrix} \overline{\sigma_n} \\ \Lambda \end{pmatrix} = \begin{pmatrix} 0 \\ f \end{pmatrix}, \qquad (8.250)$$

где T – матрица с элементами

$$t_{ij} = \int_{\Omega} \left[\frac{\partial \omega_i}{\partial x} \frac{\partial \omega_j}{\partial x} + \frac{\partial \omega_i}{\partial y} \frac{\partial \omega_j}{\partial y} \right] dxdy;$$

A – матрица, элементами которой являются

$$a_{ij} = \sum_{k=1}^{n} \omega_i(x_k, y_k);$$

f – вектор, составленный из компонент

$$f_i = \sum_{k=1}^{n} z_k \omega_i(x_k, y_k);$$

Λ – множитель Лагранжа.

В результате решения системы отыскиваются значения сплайна на подпространстве – значения функции в узлах прямоугольной сетки.

Если минимизируется функционал

$$\Phi(\Omega) = \int_{\Omega}\left[\left(\frac{\partial^2 u}{\partial x^2}\right)^2 + 2\left(\frac{\partial^2 u}{\partial x \partial y}\right)^2 + \left(\frac{\partial^2 u}{\partial y^2}\right)^2\right] dxdy, \qquad (8.251)$$

то в качестве базисных функций используются бикубические сплайны.

При отыскании сглаживающего сплайна получают систему уравнений

$(\alpha T + A)\overline{\sigma}_n = f$.

Описанный метод был разработан и реализован на ВЦ СО АН СССР [7], [25]. Аналогичный способ с использованием сплайнов на подпространстве был реализован на целочисленной арифметике при разработке автоматизированной системы крупномасштабного картографирования (АСК-1). Данный метод моделирования топографических поверхностей описывается ниже.

При построении интерполяционного сплайна на подпространстве предполагается, что получаемая поверхность H должна проходить через заданные точки

$$H(P_i) = z_i \ \ (i = 1, ..., n), \qquad (8.252)$$

и должен достигаться минимум некоторого функционала

$$\Phi(H) = \|TH\|^2 = \min. \qquad (8.253)$$

Для определенности примем, что функционал (8.253) отыскивается на множестве функций класса $w_2^1(\Omega)$

$$\Phi(H) = \iint_{\Omega}\left(\left(\frac{dH}{dx}\right)^2 + \left(\frac{dH}{dy}\right)^2\right) d\Omega = \min. \qquad (8.254)$$

С использованием метода конечных элементов из (8.252) и (8.254) можно получить [1], [2] соответственно две системы уравнений:

$$h_{ij}(i+1-x)(j+1-y) + h_{ij+1}(i+1-x)(y-j) +$$

$$h_{i+1\,j}(x-i)(j+1-y) + h_{i+1\,j+1}(x-i)(y-j) = z(x,y), \qquad (8.255)$$

где h – значения высот в узлах квадратной сетки, покрывающей всю область моделирования Ω; $z(x,y)$ – значения высот исходных точек с координатами (x,y);

$$h_{kl} - \frac{1}{9}\sum_{j=l-1}^{l+1}\sum_{i=k-1}^{k+1} h_{ij} = 0. \qquad (8.256)$$

Уравнения (8.256) справедливы только для внутренних узлов сетки квадратов, для крайних узлов они упрощаются и имеют несколько другой вид. При выводе (8.255) и (8.256) для упрощений предполагается, что сторона квадрата сетки равна единице, и координаты исходных точек также преобразованы соответствующим образом.

Очевидно, что системы (8.255) и (8.256) в общем случае будут несовместны. При использовании итерационных методов (когда для решения

каждой из систем (8.255) и (8.256) поочередно выполняется по одной итерации) получим колебательный процесс. Решение уравнений вида (8.255) дает поверхность, проходящую через заданные точки, но не очень гладкую или даже существенно не гладкую. Решив уравнения (8.256), получим поверхность более гладкую, но не проходящую через заданные точки, то есть будет нарушено условие интерполяции (8.252). После некоторого числа итераций процесс стабилизируется, и колебания будут происходить вблизи некоторого промежуточного положения с меньшей по величине, но постоянной амплитудой.

Если в процессе решения системы (8.256) при β-й итерации значения в узлах регулярной сетки вычислять по формуле

$$h_{kl}^{(\beta)} = h_{kl}^{(\beta-1)} + \alpha^{(\beta)}\left[\frac{1}{9}\sum_{j=l-1}^{l+1}\sum_{i=k-1}^{k+1} h_{ij}^{(\beta-1)} - h_{kl}^{(\beta-1)}\right], \qquad (8.257)$$

где $l = \alpha^{(1)} > \alpha^{(2)} > \ldots > \alpha^{(\gamma)} = 0$, то после окончания γ итераций полученный сплайн будет отвечать условиям (8.252) и (8.254).

Описываемый далее алгоритм является, в некоторой степени, эвристическим, поэтому остаются открытыми вопросы о существовании и единственности решения, об устойчивости и сходимости вычислительного процесса, о характере стремления α к нулю.

Прежде чем обсуждать вопрос о моделировании функций двух переменных с помощью интерполяционных сплайнов на подпространстве, рассмотрим их применение для приближения функций одной переменной.

Пусть функция $z(x)$ задана своими значениями в некоторых точках отрезка $[a,b]$ (рис. 8.109). Приближение функции одной переменной начинается с того, что область моделирования разбивается на равные конечные элементы, никак не связанные с исходными точками (рис. 8.109), и вокруг каждой исходной точки строится горизонтальный участок с высотой, равной значению функции в этой точке (рис. 8.110). Как следует из данного рисунка, приближающая функция, называемая *проксимальной*, проходит через заданные точки, но не является гладкой.

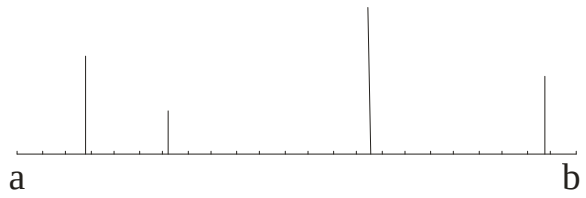


Рис. 8.109. Исходные точки и разбиение

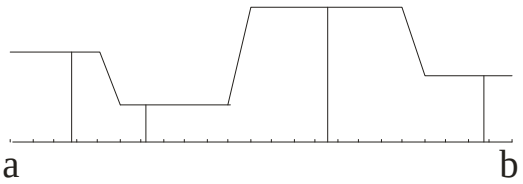


Рис. 8.110. Проксимальная функция

Процесс получения дискретной модели состоит в выполнении некоторого числа итераций. Каждая итерация состоит из процедуры сглаживания и процедуры интерполяции. Сглаживание полученной проксимальной функции осуществляется путем вычисления средних значений высот во всех внутренних узлах регулярной сетки:

$$h_i = \frac{h_{i-1} + h_{i+1}}{2},\ i = (2, ..., n-1). \qquad (8.258)$$

После выполнения сглаживания в первой итерации получим картину, представленную на рис. 8.111, на котором можно заметить тенденцию к сглаживанию при одновременном нарушении условия интерполяции. Поэтому следующей выполняется процедура интерполяции, в соответствии с которой для каждой исходной точки вычисляется поправка по формуле линейной интерполяции

$$\delta = z_k - (h_i(i+1-x_k) + h_{i+1}(i-x_k)), \qquad (8.259)$$

где $x_k \in [x_i, x_{i+1}]$, и вычисленные значения функции в двух ближайших узлах исправляются на величину δ:

$h_i = h_i + \delta$;

$h_{i+1} = h_{i+1} + \delta$.

После выполнения процедуры интерполяции функция будет проходить через заданные точки, но станет менее гладкой. На рис. 8.112 представлены исправленные значения приближающей функции в исходных точках.

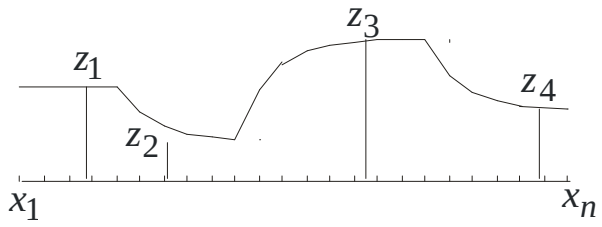


Рис. 8.111. Сглаживание в итерации 1

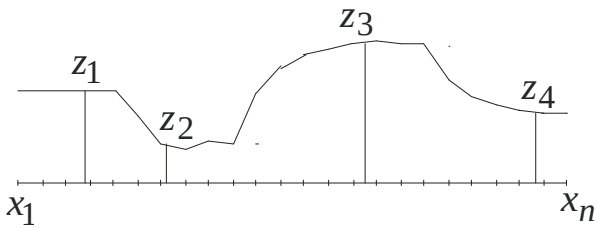


Рис. 8.112. Интерполяция в итерации 1

В следующей итерации вновь повторяются сглаживание и интерполяция, но для ускорения сходимости итерационного процесса сглаживание функции осуществляется в обратном направлении, то есть от точки x_n к точке x_1. Это означает, что система уравнений (8.256) решается в обратном направлении: от последнего уравнения к первому. Результат выполнения второй итерации после сглаживания и интерполяции представлен соответственно на рис. 8.113, 8.114. Как видим, после выполнения всего двух итераций получен не самый плохой результат. Объяснение его не только в достоинствах метода, но и в том, что в данном примере *коэффициент сгущения* – отношение числа узлов регулярной сетки к числу исходных точек – не слишком большой. При его увеличении сходимость интерполяционного процесса будет ухудшаться.

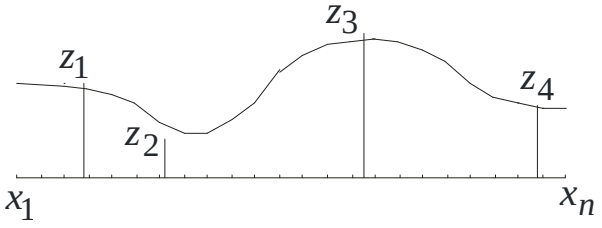


Рис. 8.113. Сглаживание

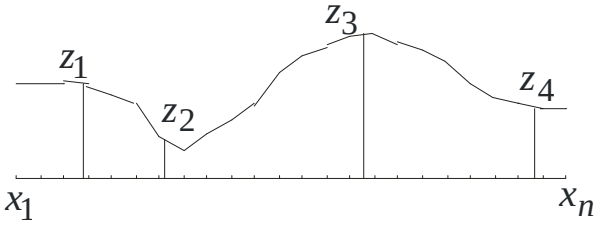


Рис. 8.114. Интерполяция

в итерации 2 в итерации 2

Рассматривая схему алгоритма построения интерполяционного сплайна на подпространстве для функции двух переменных, примем, что область моделирования Ω ограничена прямыми $x=1$, $x=m$, $y=1$ и $y=n$, покрыта сеткой квадратов со стороной, равной 1, и содержит точки $\{x_i, y_i, z_i\}$ с известными значениями функции (высотами), расположенные произвольным образом по отношению к квадратной сетке. Дополнительное ограничение – исходные точки должны располагаться внутри области Ω (рис. 8.115).

Построение интерполяционного сплайна на подпространстве осуществляется следующим образом:

1) выполняется преобразование координат исходных точек так, чтобы они отвечали указанным условиям, то есть попадали внутрь сетки квадратов;
2) определяются начальные значения функции в узлах квадратной сетки;
3) устанавливается начальное значение параметра α;
4) с помощью некоторого числа итераций уточняются значения функции в узлах сетки квадратов;
5) выполняется преобразование координат сетки в исходную систему координат.

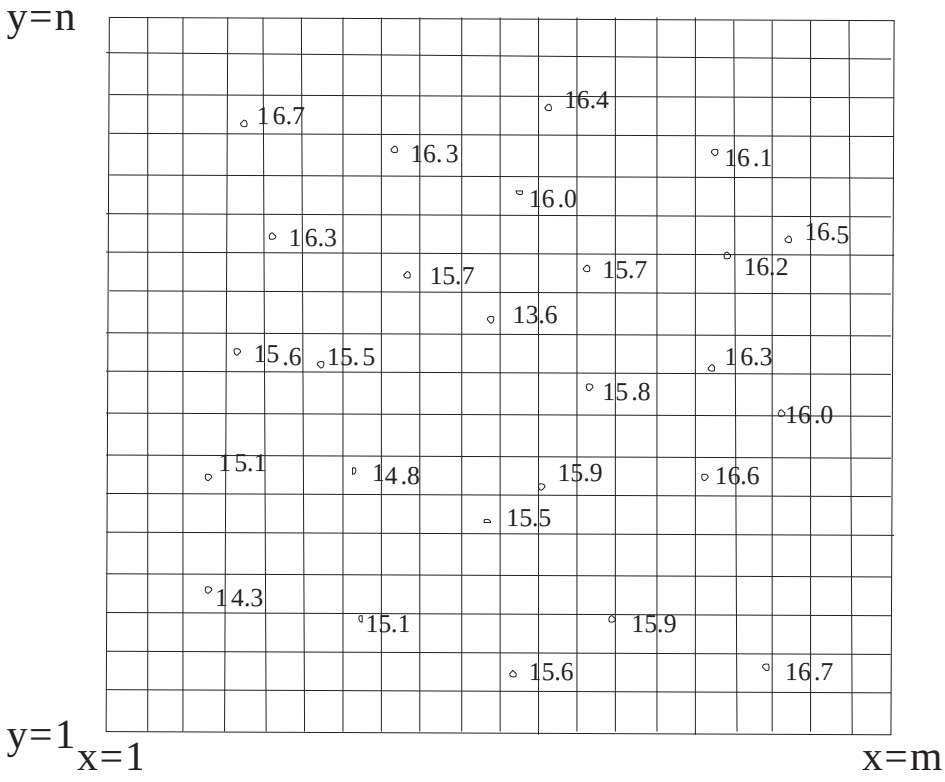


Рис. 8.115. Исходные точки и регулярная сетка

Описанная схема вычислений является наиболее общей и инвариантна по отношению к конкретным реализациям метода. Различные модификации алгоритма могут отличаться:

- способом определения начальных значений в узлах регулярной сетки;
- способом сглаживания поверхности;

– способом вычисления поправок за отклонение полученной поверхности от исходных точек;

– начальным значением, законом и скоростью изменения параметра α;

– критерием окончания итераций.

Для определения начальных значений функции в узлах модельной сетки можно использовать горизонтальную плоскость с $z = 0$ или $z = 0{,}5(z_{\min} + z_{\max})$, наклонную плоскость, полином невысокой степени, проксимальную поверхность и т. п. Наиболее эффективно использование *проксимальной поверхности*, то есть поверхности, образованной горизонтальными площадками, каждая из которых строится вокруг исходной точки P_i и описывается уравнением $H(x, y) = z_i$. Пример проксимальной поверхности приведен на рис. 8.116; для ее построения использовались данные, представленные на рис. 8.115.

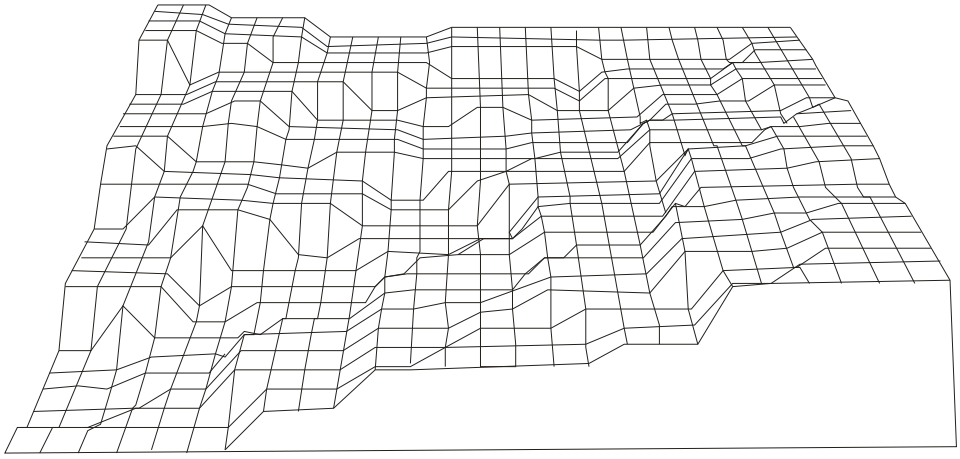


Рис. 8.116. Проксимальная поверхность

После получения приближенных значений функции в узлах регулярной сетки осуществляется их уточнение методом итераций. Каждая итерация состоит из четырех шагов:

– сглаживания поверхности;

– интерполяции – ввода поправок в значения функции в узлах сетки за отклонение полученной поверхности от исходных точек;

– вычисления нового значения параметра α;

– определения критерия окончания итераций и сравнение его с заранее указанным допуском.

Различия в способах сглаживания поверхности означают различия в способах решения системы уравнений (8.256). Сходимость итерационного процесса улучшается, если при каждой итерации используются значения неизвестных, полученные в этой же итерации, и направление решения уравнений системы (8.256) после каждой итерации меняется на противоположное.

Наиболее заметно потребность в процессорном времени уменьшается (без ощутимой потери точности), когда уравнение вида (8.256) заменяется другим, «близким» к нему уравнением. Это означает, что вместо функционала (8.254)

отыскивается другой, достаточно близкий к нему. Альтернативой выражению (8.257) может служить

$$h_{kl}^{(\beta)} = h_{kl}^{(\beta-1)} + \alpha^{(\beta)}\left(\frac{h_{k-1l}^{(\beta-1)} + h_{k+1l}^{(\beta-1)} + h_{kl-1}^{(\beta-1)} + h_{kl+1}^{(\beta-1)}}{4} - h_{kl}^{(\beta-1)}\right) \quad .$$

(8.260)

Иными словами, вместо восьми близлежащих узлов используются только четыре (рис. 8.117).

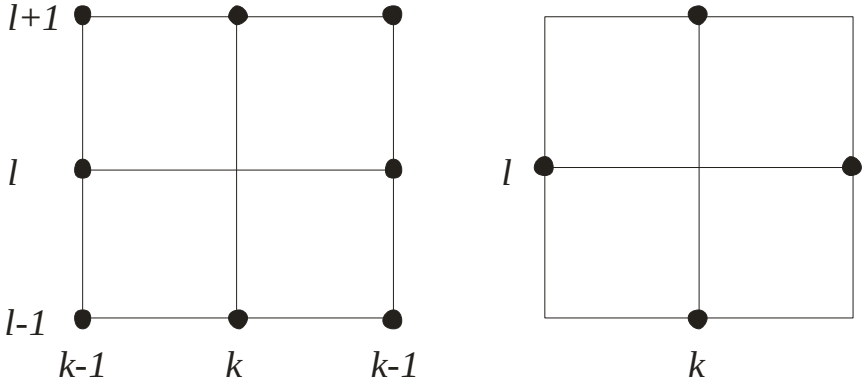


Рис. 8.117. Исключение узлов

Более того, функционал (8.254) может быть заменен двумя функционалами, каждый из которых отыскивается на множестве функций класса $w_1^1(\Omega)$. Такое решение приводит к замене системы (8.257) двумя системами уравнений вида

$$h_{kl}^{(\beta)} = h_{kl}^{(\beta-1)} + \alpha^{(\beta)}\left(\frac{h_{k-1l}^{(\beta-1)} + h_{k+1l}^{(\beta-1)}}{2} - h_{kl}^{(\beta-1)}\right); \qquad (8.261)$$

$$h_{kl}^{(\beta)} = h_{kl}^{(\beta-1)} + \alpha^{(\beta)}\left(\frac{h_{kl-1}^{(\beta-1)} + h_{kl+1}^{(\beta-1)}}{2} - h_{kl}^{(\beta-1)}\right). \qquad (8.262)$$

Далее возможны два варианта: либо в процессе выполнения каждой итерации решаются обе системы (8.261) и (8.262), либо при нечетной итерации решается одна из них, а при четной – другая. В первом случае время выполнения каждой итерации больше, но требуется их меньшее число, во втором – наоборот. Результаты экспериментов показали, что первый вариант более эффективен.

С возрастанием *коэффициента сгущения* уменьшается скорость итерационного процесса, но может быть повышена точность моделирования. Кроме того, при увеличении числа узлов сетки квадратов линейно возрастает время выполнения одной итерации. Эффективным средством ускорения сходимости может служить разбиение итерационного процесса на несколько этапов (2–3) с уменьшением шага сетки в два раза при переходе от этапа к этапу.

Неоднозначное решение уравнений (8.255) следует из того, что в каждое из них входит четыре неизвестных. Неоднозначность решения (8.255) можно устранить введением *принципа равенства поправок* в значения неизвестных

$$\left.\begin{array}{l} h_{ij}^{(\beta)} = h_{ij}^{(\beta-1)} + \delta^{(\beta)} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ h_{i+1\,j+1}^{(\beta)} = h_{i+1\,j+1}^{(\beta-1)} + \delta^{(\beta)} \end{array}\right\}, \qquad (8.263)$$

что равносильно параллельному смещению элемента поверхности по вертикали. Значение поправки δ вычисляется по формуле

$$\delta = z(x,y) - [h_{ij}(i+1-x) + h_{i+1\,j}(x-i)](j+1-y) +$$

$$+ [h_{ij+1}(i+1-x) + h_{i+1\,j+1}(x-i)](y-j). \qquad (8.264)$$

Из формулы (8.264) следует, что если на конечный элемент приходится несколько исходных точек, то в общем случае невязки (отклонения) могут быть устранены только при условии уменьшения размеров конечного элемента.

Другой возможностью для вычисления исправленных значений функции в узлах сетки является использование выражений

$$\left.\begin{array}{l} h_{ij}^{(\beta)} = h_{ij}^{(\beta-1)} + \delta^{(\beta)}(i+1-x)(j+1-y) \\ \ldots \\ h_{i+1\,j+1}^{(\beta)} = h_{i+1\,j+1}^{(\beta-1)} + \delta^{(\beta)}(x-i)(y-j) \end{array}\right\}, \qquad (8.265)$$

то есть применение сглаживающего сплайна на подпространстве [21].

При малом значении интервала между узлами сетки и сравнимыми с ним ошибками определения планового положения исходных точек можно использовать выражение

$$\delta = z(x,y) - \frac{1}{4}(h_{ij} + h_{i+1\,j} + h_{ij+1} + h_{i+1\,j+1}) \qquad (8.266)$$

или даже

$$\delta = z(x,y) - h_{ij}, \qquad (8.267)$$

но приближение будет более грубым.

В качестве *начального значения* α выбирается 1. Присваивание α большего начального значения (хотя бы 1.5) может привести к тому, что в некоторых случаях итерационный процесс начнет расходиться. При начальном значении $\alpha^{(1)} < 1$ скорость достижения минимума функционала (8.254) падает и требуется большее число итераций.

При переходе от итерации к итерации для вычисления α может использоваться выражение

$$\alpha^{(\beta)} = \alpha^{(\beta-1)} - d_\alpha \ (0 < d_\alpha << 1), \qquad (8.268)$$

более предпочтительное, чем

$$\alpha^{(\beta)} = c\alpha^{(\beta-1)} \ (0 < c < 1). \qquad (8.269)$$

Из физических соображений следует, что скорость убывания должна быть невысокой, по крайней мере, на последних итерациях. Поэтому α можно представить кубическим полиномом от числа итераций, но тогда потребуется задавать число итераций, что нежелательно.

В качестве *критерия окончания итераций* могут быть приняты:

– среднеквадратическое или наибольшее по модулю (как более определенное) отклонение полученной поверхности от исходных точек;

– среднеквадратическая или наибольшая по модулю поправка за кривизну поверхности;

– абсолютное или относительное изменение функционала (8.254) между двумя итерациями (не обязательно соседними);

– разности между отклонениями поверхности от исходных точек из двух итераций и другие.

Данный метод моделирования обладает следующими свойствами:

– алгоритмичностью, модифицируемостью и наглядной интерпретацией;

– линейной зависимостью времени t выполнения одной итерации от числа исходных точек и числа узлов модельной сетки

$$t = a + bk + cmn,$$

где a, b, c – коэффициенты, зависящие от типа процессора и реализации алгоритма; k – число исходных точек; m, n – число узлов регулярной модели по осям x и y;

– устойчивостью к ошибкам в исходных данных, их влияние ограничено ближайшими исходными точками;

– отсутствием потребности в большом объеме оперативной памяти для хранения коэффициентов систем линейных уравнений.

Кроме того, он обладает возможностью:

– применения для сглаживания и для интерполяции;

– распространения на случай функций трех переменных;

– получения информационных моделей различной точности и степени обобщения;

– моделирования топографической поверхности по частям в условиях недостаточного объема оперативной памяти;

– использования в качестве начального приближения полученной ранее модели, если она признана неудовлетворительной;

– наложения «заплат» на модель при обнаружении ошибок в исходных данных или при старении модели;

– реализации на целочисленной арифметике;

– высокого распараллеливания алгоритма;

– реализации бикубических сплайнов на подпространстве и алгоритмов, использующих веса исходных точек.

Достоинством описанного метода является устойчивость к ошибкам в исходных данных. При моделировании одного из участков со спокойным рельефом, на котором колебания высот не превышали 100 м, а превышения

между соседними точками были не больше 10 м, была допущена непреднамеренная ошибка: высота одной из точек была больше фактической более чем на 800 м. Программа справилась с поставленной задачей. Полученная поверхность проходила через заданные точки, а в точке с ошибочной отметкой был создан столб высотой 800 м. Горизонтали были неправильно отрисованы только на ребрах, соединяющих ошибочную точку с соседними, на остальную область влияние ошибки не распространилось.

Данный метод был реализован в автоматизированной системе картографирования АСК-1 на ЕС ЭВМ. Для представления координат, высот исходных точек и высот в узлах модельной сетки использовались целые числа длиной два байта. Число узлов регулярной сетки могло достигать 360 000. Если средние размеры трапеции (планшета) принять 60×60 см (считая полосу перекрытия с соседними участками для обеспечения сводки горизонталей), то при этих условиях величина интервала между узлами регулярной сетки будет составлять 1 мм. Очевидно, что для многих приложений такое разрешение можно признать достаточным.

В качестве примера возможностей метода и возможности его использования в сочетании со структурными линиями топографической поверхности рассмотрим участок поверхности с оврагом, представленный на рис. 8.118. Для построения модели были отобраны точки на горизонталях и на структурных линиях (верх и низ оврага), в совокупности образующих каркас поверхности. Перед моделированием выполнялось сгущение точек на структурных линиях. На рис. 8.119 представлена аксонометрическая проекция поверхности, полученной в результате моделирования. Шаг модельной сетки составлял 0.25 мм, поэтому на рис. 8.119 она изображена с разрежением. Качество полученной модели можно признать удовлетворительным. Однако, необходимо подчеркнуть, что для правильного отображения локальных особенностей потребовались представление структурных линий и высокая разрешающая способность модели.

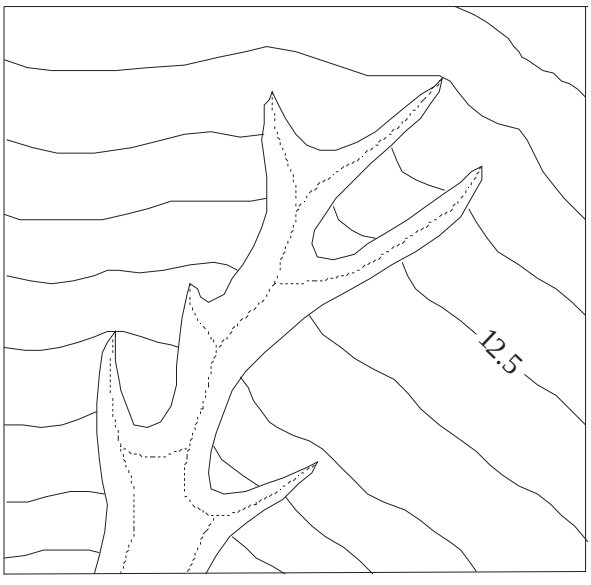


Рис. 8.118. Участок с оврагом

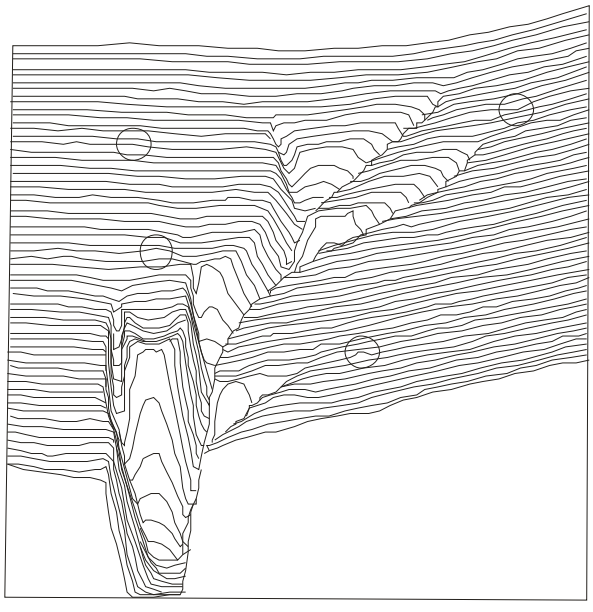


Рис. 8.119. Аксонометрическая проекция

Поверхности, получаемые данным методом, очень близки к минимальным поверхностям. *Минимальной поверхностью* называют поверхность, средняя кривизна которой во всех точках равна нулю. Математической записью данного свойства является выражение (8.254). Физической моделью минимальной поверхности может служить мыльная (или резиновая) пленка, натянутая на проволочный каркас, представляющий собой замкнутую пространственную кривую. В случае плоской кривой это будет кусок плоскости, ограниченный данной кривой. Данное объяснение минимальной поверхности было предложено в 1849 г. Ж. Плато.

В нашем случае к пространственному каркасу из проволоки внутри области моделирования добавляются точки разной высоты, в которых пленка закрепляется. На рис. 8.119 подобные точки просматриваются, и некоторые из них отмечены кружком. При уменьшении шага модельной сетки получаемая поверхность будет приближаться к минимальной поверхности.

В целом о моделировании топографических поверхностей с применением сплайнов на подпространстве можно сказать, что этот метод, возможно, не самый красивый и не самый лучший, но довольно простой и безопасный. Однако, необходимо еще раз подчеркнуть, что для правильного отображения локальных особенностей требуются представление структурных линий и высокая разрешающая способность модели. Интересной особенностью описанного метода является возможность его использования в модифицированном виде для моделирования процессов развития рельефа во времени.

Способы непосредственного пересчета сетки с хаотично расположенными узлами на регулярную сетку, основанные на методе конечных элементов, отличаются вычислительной эффективностью, умеренными требованиями к объему оперативной памяти, хорошим качеством приближения, хотя и требуют высокой разрешающей способности регулярной дискретной модели. Пожалуй, наиболее ценными свойствами рассмотренных методов этого класса являются возможность очень высокого распараллеливания операций и простота, алгоритмичность. Возможно, что имеется даже смысл в аппаратной реализации подобных методов, поскольку задача приближения функций, заданных своими эмпирическими значениями, встречается довольно часто в самых разнообразных приложениях.

8.28. Отображения непрерывного множества на дискретное

Отображения непрерывного множества на дискретное не играют такой роли, как отображения двух классов, рассматривавшихся выше, хотя используются чаще. Значение рассмотренных выше методов объясняется тем, что результаты любого метода сбора данных о топографических поверхностях представляют собой дискретные множества. Таким образом, эти способы в системах моделирования топографических поверхностей используются почти всегда. Совсем иначе обстоит дело с аналитическими моделями. Они могут быть получены следующим образом:

– либо в процессе проектирования, когда поверхность или совокупность поверхностей описывается уравнением (уравнениями);

– либо в результате работы системы моделирования как вторичная модель.

Представление непрерывного множества дискретным – это задача нахождения значений функции одной или двух переменных на некоторой системе точек. На рис. 8.120 изображена функция одной переменной, аналитическое представление которой известно. В отличие от рассматривавшихся ранее задач эту функцию нужно заменить некоторым множеством точек таким образом, чтобы ошибка, например, линейной интерполяции между ними, не превышала некоторого допуска. Данную задачу можно усложнить и потребовать, чтобы число таких точек было минимальным.

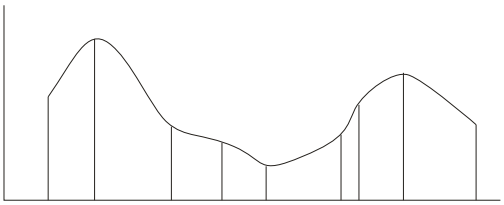


Рис. 8.120. Определение характерных точек

Хотя тут возможны свои сложности, например, вычисление значений функции, не выражающейся в явном виде, из математического анализа известны разнообразные способы их преодоления, и техника получения таких решений хорошо разработана. Задача заметно усложняется, если система точек не задана, а должна быть определена так, чтобы достигался некоторый оптимум, например, минимизировалось число узлов при заданном значении наибольшего отклонения между узлами от линейной интерполяции. Принципы решения задач такого рода могут быть найдены в математической литературе. Кроме того, потребность в подобных методах при геомоделировании практически не возникает и поэтому здесь они не рассматриваются. Сюда же может быть отнесена задача восстановления каркаса – структурных линий и точек топографической поверхности – по аналитическому выражению. Анализ выражения и автоматическое определение некоторых структурных точек (вершин, котловин и седловых точек) является решаемой проблемой. Для этого необходимо найти все точки, в которых первые производные равны 0. Однако решение данной задачи в полном объеме характеризуется крайней сложностью, и такие попытки, вероятно, вообще не предпринимались.

8.29. Отображения непрерывного множества на непрерывное

Отображение непрерывного множества на непрерывное – это замена одного аналитического выражения другим, которое может найти применение в связи с приближенными расчетами на топографической поверхности. Вообще же, маловероятно, что после получения точного аналитического представления поверхности будет создаваться его огрубленная модель, которая и будет использоваться в дальнейшем. Если такого рода преобразования будут использоваться при моделировании топографических поверхностей, то только при условии, что дополнительные затраты на переработку модели окупятся при

решении задач пользователей, и точность решений окажется вполне удовлетворительной.

Побудительным мотивом к замене одной аналитической модели другой аналитической моделью может служить очень большое число вычислений при представлении поверхности одним уравнением, особенно если в нем присутствуют математические функции. Обычным решением является отбрасывание последних членов описывающего поверхность ряда, если он достаточно быстро убывает. Известны также способы, когда один ряд заменяется другим с большей скоростью убывания его членов.

Минимизация максимальной ошибки аппроксимации выполняется с использованием полиномов Чебышева. На них основано также свертывание степенных рядов или оптимизация Ланцоша. При аппроксимации по методу наименьших квадратов в общем случае приходится решать систему нормальных уравнений. Использование многочленов Лежандра дает возможность вычислять коэффициенты непосредственно, без решения системы уравнений [23].

Изучение методов данного класса в применении к моделированию топографических поверхностей представляет скорее академический интерес, чем практический. Во-первых, при их использовании может происходить заметное изменение планового положения экстремальных точек, что, безусловно, вызовет возражения со стороны пользователей. Во-вторых, непрерывные информационные модели не способны и едва ли будут способны составить конкуренцию кусочно-непрерывным регулярным и нерегулярным моделям. При получении аналитического представления поверхностей по регулярным моделям используются простые уравнения, и потребность в использовании методов данного класса характеризуется чуть ли не нулевой вероятностью. В-третьих, если бы в качестве стандарта были выбраны аналитические модели, то по понятным причинам пришлось бы резко ограничить класс используемых функций.

Не хотелось бы утверждать, по крайней мере – в категоричной форме, невозможность или нецелесообразность использования композиции методов. Напротив, при разработке систем информационного моделирования топографических поверхностей допустимо любое сочетание методов, обеспечивающее необходимую точность и высокую эффективность. Однако, представляется сомнительным, что композиция более чем двух методов способна обеспечить приемлемые результаты, так как удлинение цепи преобразований неизбежно связано с увеличением времени вычислений и, кроме того, любое преобразование информации сопровождается ее искажениями. Вероятнее всего, наибольшее распространение получат методы непосредственного пересчета произвольной сетки в регулярную, подобные описанному выше методу интерполяционных сплайнов на подпространстве.

Анализ различных методов моделирования неизбежно носит качественный, оценочный характер. Более или менее уверенно можно говорить о точности и объеме оперативной памяти. Что касается времени вычислений (и стоимости), то оценки могут быть только весьма приближенные. Непосредственно можно сравнить только процессорное время. Но, поступая

таким образом, мы выносим суждение не о самом методе, а о его реализации, поскольку даже при нескольких реализациях одного метода время счета на ЭВМ может существенно различаться. Поэтому сравнение вычислительной эффективности разных методов весьма проблематично. Непосредственно в математическом методе заложены лишь потенциальные возможности, их реализация может потребовать значительных усилий.

Более того, оценка относится даже не столько к методу информационного моделирования, сколько к функционированию системы моделирования топографических поверхностей в целом. Очевидно, что ее эффективность будет зависеть от структуры исходных данных, структуры модели и архитектуры всего программного комплекса.

В реальных условиях точность и стоимость информационных моделей являются функциями от большого числа переменных: размеров участка моделирования, сложности поверхности, метода сбора данных, используемого при сборе оборудования, квалификации исполнителей, критерия выборки исходных точек, их плотности, используемого метода моделирования, характеристик компьютера (частоты процессора, объема оперативной памяти, периферийных устройств), сложившегося уровня цен на средства вычислительной техники, организации работ в предприятии и некоторых других. Исследование всего многообразия факторов и их взаимосвязей не представляет легкой задачи даже для отдельной организации.

Но нет особой нужды доказывать, что решающее значение принадлежит математическому обеспечению, то есть ядру системы информационного моделирования топографических поверхностей. Стоимость разработки программных средств сегодня намного превышает уровень затрат на приобретение аппаратных средств, и наблюдается устойчивая тенденция к увеличению этого разрыва. Разработка информационных моделей и методов моделирования играет важную роль в повышении эффективности геомоделирования, поэтому исследования в данной области продолжаются.

8.30. Создание горизонталей по сетке квадратов

Создание горизонталей по информационной модели топографической поверхности относится к числу обязательных функций автоматизированной картографической системы, поэтому ее реализация требует тщательной проработки. От качества решения данной задачи может существенно зависеть как время их построения в автоматическом режиме, так и объем доработок картографического изображения в интерактивном режиме. Описание алгоритмов построения горизонталей здесь также приводится как пример решения одной из прикладных задач с использованием информационной модели топографической поверхности и пример задачи, содержащей элементы нечисленного программирования.

Известны два основных алгоритма отслеживания изолиний на прямоугольной сетке. В каждом из них горизонтали отслеживаются последовательно одна за другой. Общая схема обоих алгоритмов при отслеживании горизонтали с высотой H делится на два этапа:

1. отслеживание разомкнутых горизонталей;
2. отслеживание замкнутых горизонталей.

Отслеживание разомкнутой горизонтали начинается с поиска ее начала. Для этого просматриваются все внешние ребра модельной сетки и проверяется условие

$$(H_1 - H)(H_2 - H) < 0, \qquad (8.270)$$

где H_1 и H_2 – значения высот в двух соседних узлах сетки; H – значение высоты отслеживаемой горизонтали. Если условие не выполняется, то переходят к следующему ребру сетки, пока не будут просмотрены все ее внешние ребра, либо не найдено ребро, на котором это условие выполняется. Как только такое ребро обнаружено (говорят, что горизонталь «вошла в квадрат»), осуществляется ее отслеживание:

– находят точку «выхода» горизонтали из квадрата;

– поскольку точка выхода из одного квадрата является точкой входа в другой квадрат, то отыскивается точка выхода из следующего квадрата и т. д., пока горизонталь не выйдет на внешнее ребро сетки, что является признаком конца данной горизонтали.

После этого продолжают просмотр оставшихся внешних ребер, так как в пределах одного участка моделирования может быть несколько разомкнутых горизонталей с одной высотой и их число заранее не известно. Как только все разомкнутые горизонтали с заданной высотой отслежены, приступают к поиску замкнутых горизонталей с той же высотой. Для этого последовательно просматриваются все параллельные либо оси *x*, либо оси *y* внутренние ребра сетки. Это следует из того, что любая замкнутая горизонталь пересекает, как минимум, два горизонтальных и два вертикальных ребра сетки. Поэтому для ее обнаружения достаточно проверки всех ребер какого-либо одного направления. После обнаружения замкнутой горизонтали ее прослеживание выполняется так же, как и разомкнутой, с той лишь разницей, что признаком окончания горизонтали является возвращение в ее начальную точку.

В процессе поиска разомкнутой или замкнутой горизонтали неизбежно возникает ситуация, когда просматривается ребро, через которое горизонталь уже прошла. Чтобы не отслеживать горизонталь несколько раз, такие ребра должны каким-то образом помечаться. Известные алгоритмы различаются способом решения этой задачи.

В первом из них используется вспомогательный двумерный массив, каждый элемент которого соответствует узлу сетки квадратов. Перед началом отслеживания горизонталей с высотой H всем элементам вспомогательного массива присваивается значение 0.

Пусть, для определенности, поиск начала изолинии осуществляется проверкой ребер, параллельных оси *x*. Если в процессе прослеживания горизонталь пересекает ребро с высотами H_{ij} и $H_{i+1\,j}$, то элементу вспомогательного массива, соответствующему узлу (x_i, y_j), присваивается значение, отличное от 0. В дальнейшем при поиске начала замкнутой горизонтали

такие ребра пропускаются. Перед отслеживанием горизонтали с другой высотой всем элементам вспомогательного массива присваивается значение 0 и процесс повторяется. Недостатком описанного алгоритма является потребность в значительном объеме оперативной памяти для хранения вспомогательного массива.

Во втором алгоритме вспомогательный массив не требуется. Если в некоторых узлах сетки значения высоты $H_{ij} \le 0$, то все высоты смещаются на одно и то же значение таким образом, чтобы все новые значения высот были положительными. Иными словами, вводится новая система высот.

Если в процессе отслеживания разомкнутая или замкнутая горизонталь пересекает некоторое ребро, параллельное оси *x*, то значению высоты в левом узле ребра присваивается знак «минус». В дальнейшем при поиске начала замкнутой горизонтали такие ребра пропускаются. Но, как и в первом алгоритме, каждый узел просматривается *дважды*:

– перед началом поиска горизонтали с высотой *H* все узлы проверяются, и если значение высоты в узле меньше нуля, ему присваивается знак «плюс»;

– выполняется поиск горизонтали, для чего просматриваются все внешние ребра и все внутренние ребра, параллельные оси *x*.

В описанном ниже алгоритме отслеживание всех горизонталей с фиксированной высотой *H* выполняется *за один просмотр* сетки квадратов.

Пусть, как и во втором алгоритме, все высоты изменяются на некоторую константу так, чтобы выполнялось соотношение

$$H_{ij} > 0 \ (i = 1, ..., m;\ j = 1, ... n).$$

Вначале отслеживаются все разомкнутые горизонтали с высотой *H*, затем все замкнутые горизонтали с той же высотой, после чего осуществляется переход к следующей горизонтали и т. д. С некоторыми упрощениями алгоритм описывается ниже.

В алгоритме построения горизонталей по регулярной модели на сетке квадратов можно выделить четыре этапа:

1) поиск разомкнутой горизонтали;
2) отслеживание разомкнутой горизонтали;
3) поиск замкнутой горизонтали;
4) отслеживание замкнутой горизонтали.

Поиск и отслеживание разомкнутой горизонтали разбивается на этапы:

– поиск по левой границе;
– поиск по верхней границе;
– поиск по правой границе;
– поиск по нижней границе.

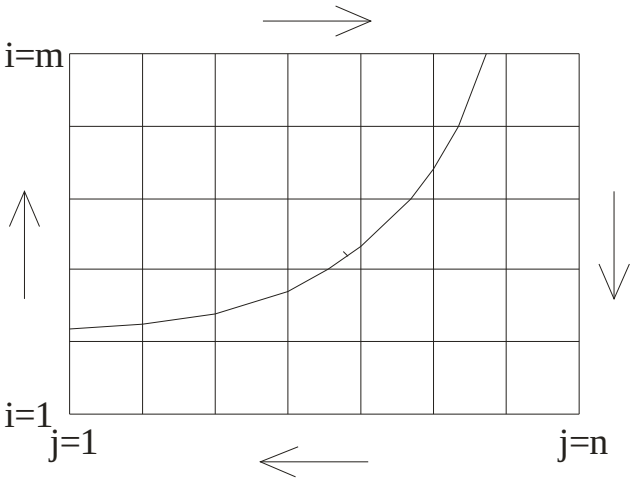


Рис. 8.121. Направление обхода

В алгоритме поиска разомкнутых горизонталей важны два момента. Просмотр внешних ребер сетки осуществляется по часовой стрелке (рис. 8.121). Горизонталь обнаруживается только тогда, когда высота текущего узла больше или равна высоте горизонтали *H*, а значение высоты следующего по направлению обхода границы узла меньше *H*. В результате каждая разомкнутая горизонталь будет обнаружена только в точке входа. В точке входа поверхность имеет уклон влево (если двигаться по горизонтали от границы сетки в ее внутреннюю область), а в точке выхода – уклон вправо (если опять двигаться от границы вдоль горизонтали, то есть в обратном направлении). Следовательно, ни одна разомкнутая горизонталь не может быть обнаружена дважды.

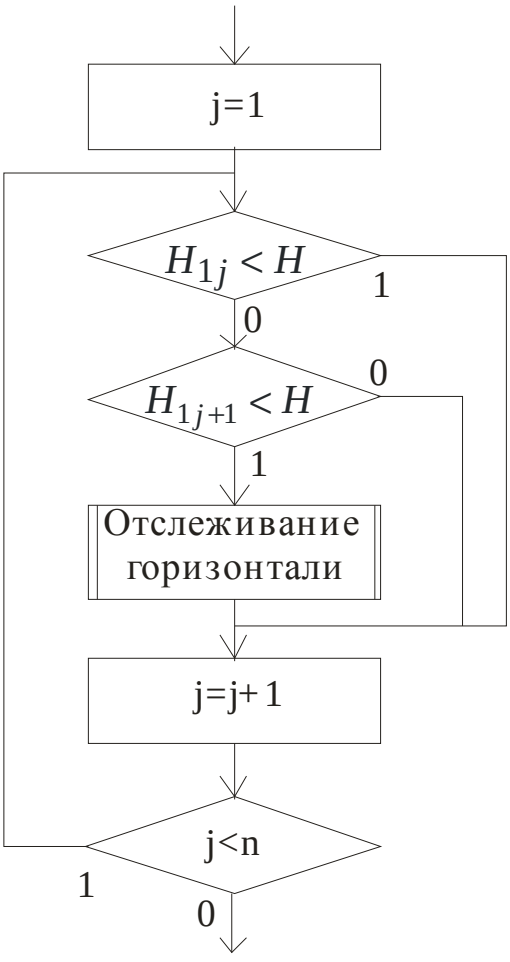


Рис. 8.122. Поиск разомкнутой горизонтали. В. 1

На рис. 8.122 представлена блок-схема двух алгоритмов поиска горизонтали по левому краю сетки квадратов. На этом рисунке H_{ij} обозначает высоту соответствующего узла, а *H* – высоту горизонтали. Этот алгоритм более понятен, но можно заметить, что значения высот многих узлов будут проверяться дважды: вначале как H_{1j+1}, а затем как H_{1j}. От этого недостатка избавлен алгоритм на рис. 8.123, возможно, менее понятный, но более быстрый.

Поиск по верхней, правой и нижней границам прямоугольной сетки осуществляется аналогичным образом с небольшими очевидными поправками. После того, как разомкнутая горизонталь обнаружена, начинается ее отслеживание.

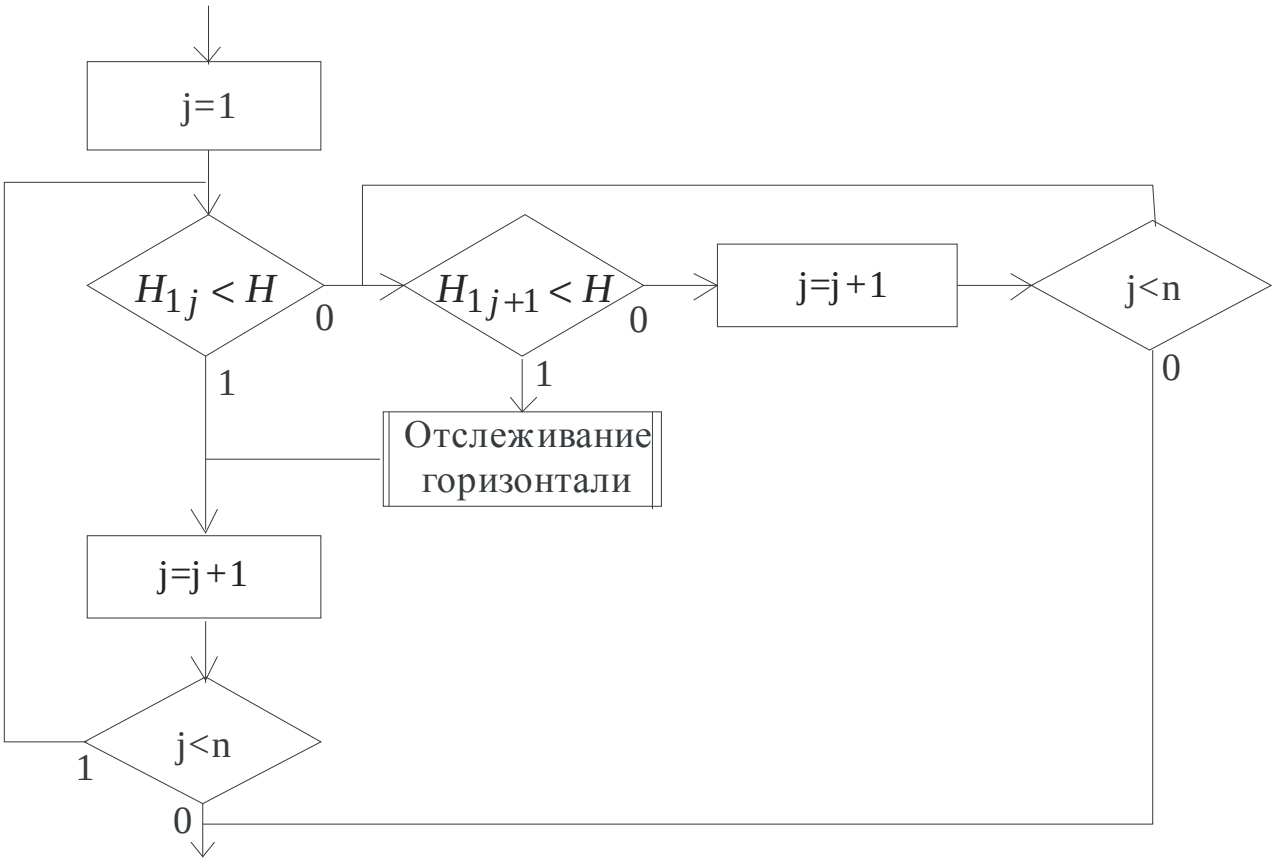


Рис. 8.123. Поиск разомкнутой горизонтали. В 2

Алгоритм отслеживания разомкнутой горизонтали представлен на рис. 8.124. Входными параметрами для процедуры отслеживания являются индексы i и j точки левого нижнего квадрата, в который вошла горизонталь, а также признак стороны p, с которой горизонталь вошла в квадрат (его значения даны на рис. 8.124). Переменная q является внутренней и указывает на сторону выхода горизонтали из квадрата. Отслеживание горизонтали заканчивается, когда она выходит на границу сетки квадратов, то есть когда выполняется условие

$$(i = 1) \vee (i = m) \vee (j = 1) \vee (j = n).$$

Здесь необходимо обратить внимание на то, что помечаются только те горизонтальные ребра сетки (присваивается знак «минус» их левому узлу), которые пересекаются изолинией сверху вниз (с севера на юг). Тогда отрицательные значения высот могут иметь только некоторые узлы на левой границе сетки или некоторые внутренние узлы (на рис. 8.125 они помечены). Необходимость такой пометки узлов будет объяснена дальше.

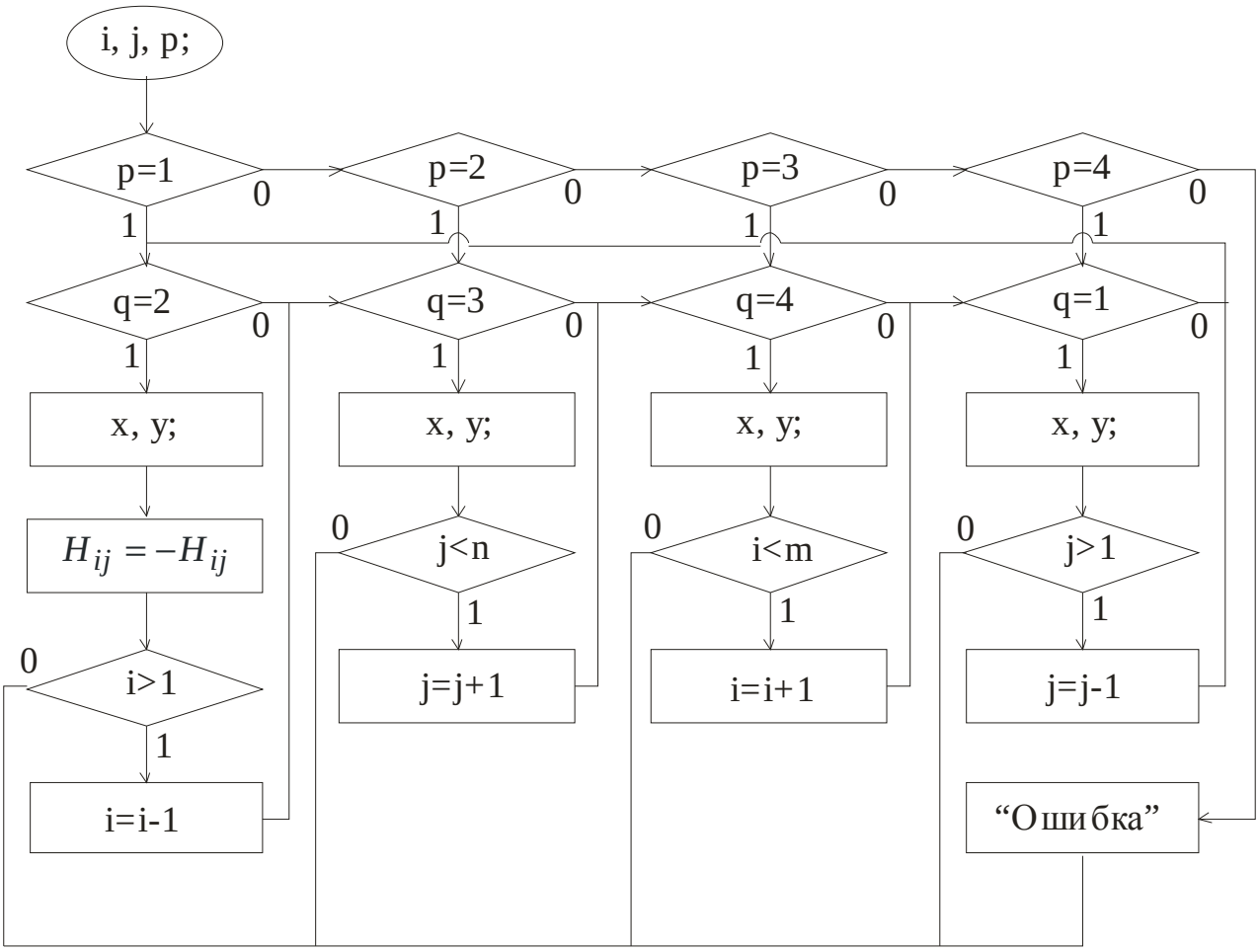


Рис. 8. 124. Отслеживание разомкнутой горизонтали:
p = 1 – вход слева; q = 1 – выход слева;
p = 2 – вход снизу; q = 2 – выход снизу;
p = 3 – вход справа; q = 3 – выход справа;
p = 4 – вход сверху; q = 4 – выход сверху

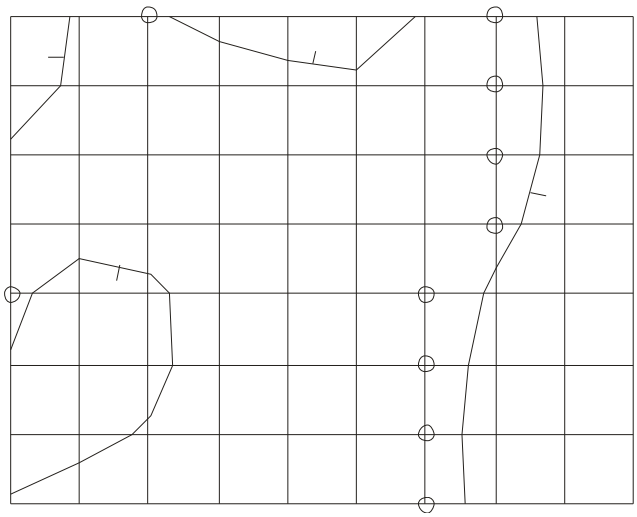
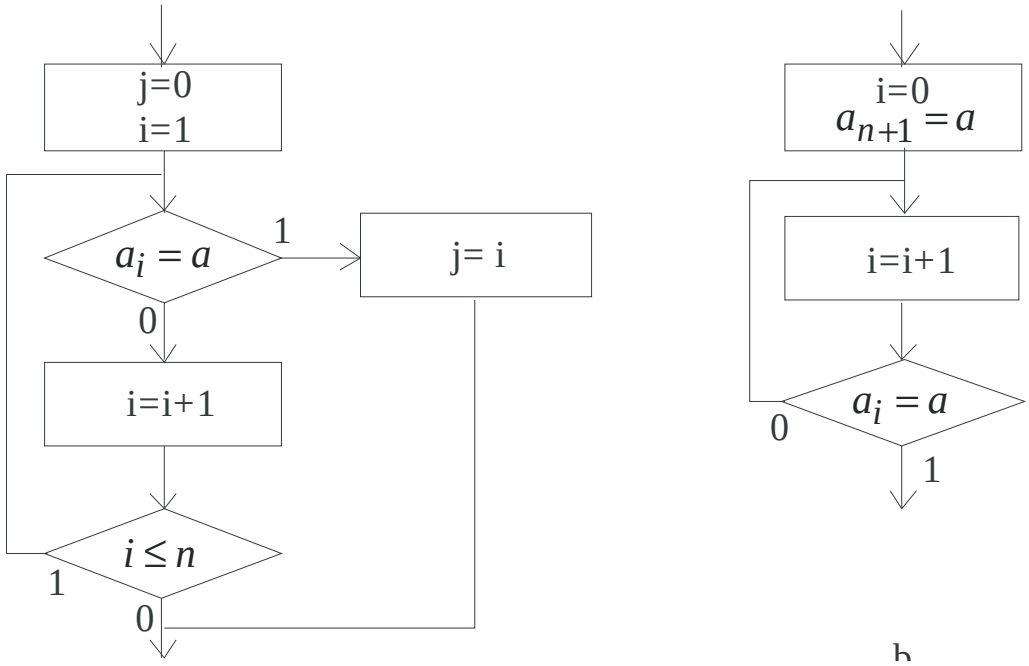


Рис. 8.125. Пересечение горизонтальных ребер

Прежде чем перейти к изложению алгоритмов поиска и прослеживания замкнутых горизонталей, опишем один прием, который применяется при поиске некоторого элемента в неупорядоченном массиве и в измененном виде используется ниже. Прием состоит в следующем. Пусть требуется ответить на вопрос, есть ли среди всех *n* элементов массива *A* хотя бы один, имеющий значение *a*. Число элементов массива увеличивается на 1, и этому

новому элементу a_{n+1} присваивается значение a. Следовательно, элемент будет найден всегда. Если его порядковый номер $i < n+1$, то это означает, что элемент в массиве есть. Если $i = n+1$, то такого значения в первоначальном массиве не было. Этот прием позволяет значительно сократить число операций за счет того, что в цикле не нужно выполнять проверку на окончание массива.

На рис. 8.126 представлены два варианта алгоритма поиска элемента в неупорядоченном массиве: слева – без добавления элемента в массив, справа – с добавлением элемента. Очевидно, что второй вариант намного проще и быстрее.



а) б)

Рис. 8. 126. Поиск элемента в массиве

Пользуясь случаем, можно выразить восхищение программистом, первым задумавшимся над решением этой задачи. Она представляется настолько простой, что, скорее всего, подавляющее большинство программистов, не задумываясь, реализовали бы первый вариант алгоритма.

Поскольку узлы правой границы сетки не могут принимать отрицательные значения, то поиск начала замкнутой горизонтали может быть организован следующим образом (рис. 8.127).

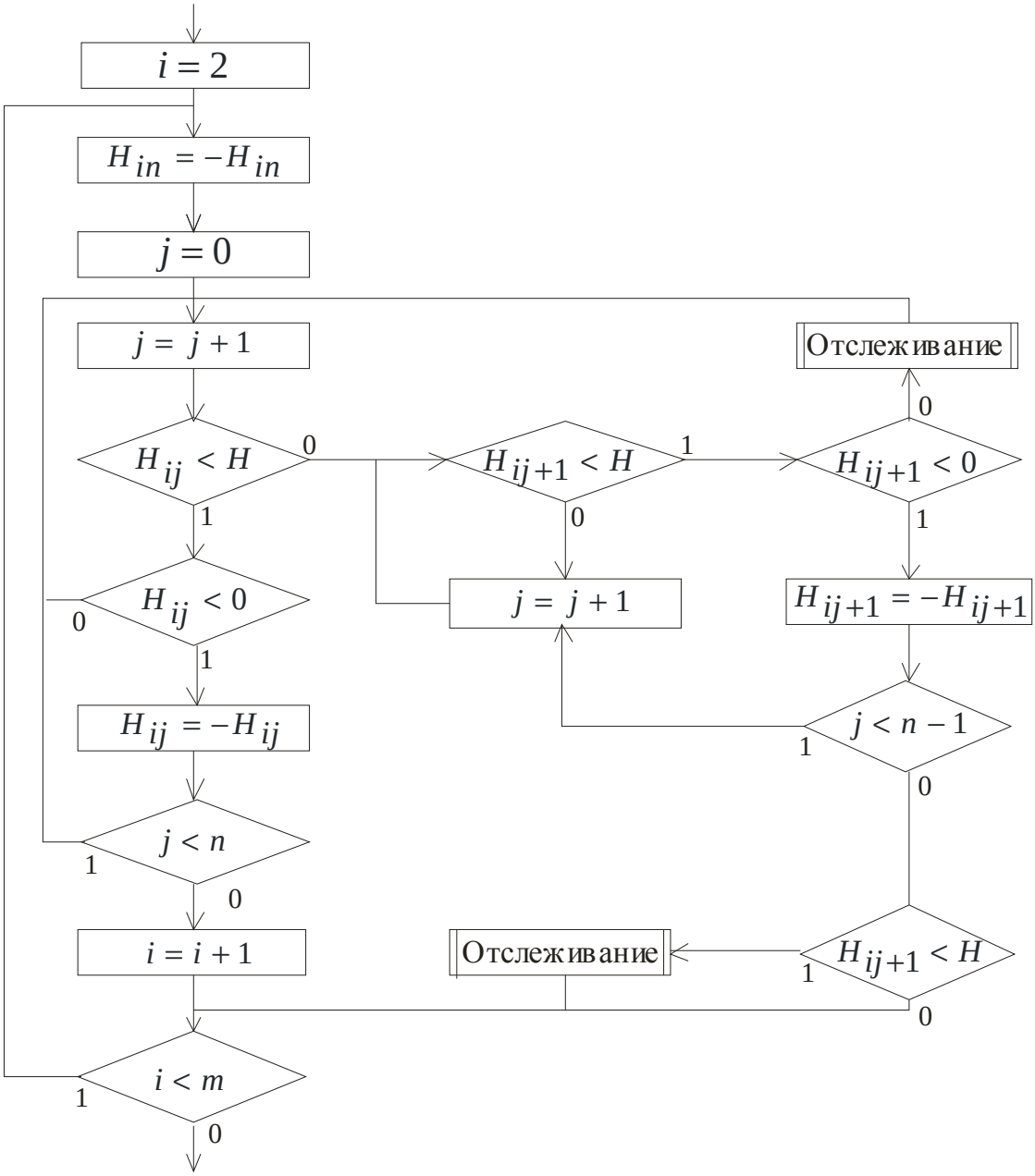


Рис. 8. 127. Поиск замкнутой горизонтали:
H_{ij} – значение высоты в узле;
$1 \le i \le m$ – индекс строки узлов;
$1 \le j \le n$ – индекс узла в строке

Прослеживание замкнутых горизонталей в общих чертах выполняется таким же образом, как и прослеживание разомкнутых. Отличие состоит в следующем:

– обход горизонтали осуществляется так, что поверхность при этом всегда имеет уклон влево (на рис. 8.128 начальная точка горизонтали отмечена кружком, а направление обхода – стрелкой);

– помечаются только те горизонтальные ребра сетки, которые изолинией пересекаются «сверху вниз» (на рис. 8.128 левые узлы этих ребер помечены точками);

– признаком окончания является отрицательное значение высоты в левом узле ребра сетки (возвращение в начальную точку);

– при возвращении горизонтали в начальную точку восстанавливается значение высоты в левом узле – значение меняется на противоположное.

Основная часть процессорного времени в алгоритме построения горизонталей приходится на этап поиска начала замкнутых горизонталей, поскольку эта операция требует примерно в *n* раз больше времени, чем поиск начала разомкнутой горизонтали. Кроме того, при увеличении числа узлов квадратной сетки возрастает и относительное время любого способа поиска замкнутых горизонталей. В рассмотренном способе это время минимально по сравнению с другими алгоритмами.

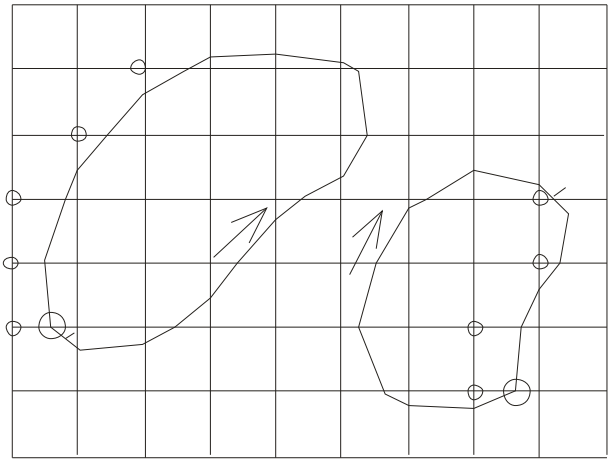


Рис. 8. 128. Начальные точки и направления обхода замкнутых горизонталей

Эффект снижения вычислительных затрат при создании горизонталей достигается в результате:

– использования оптимальной общей структуры алгоритма;

– реализации метода с использованием представления арифметических данных как целых чисел длиной 2 байта, то есть дискретизации высот;

– реализации алгоритма на языке ассемблера;

– предположением, что сетка квадратов является единичной;

– использованием простого теста на прохождение горизонтали через ребро;

– позволяет избавиться от лишней проверки (после просмотра каждого ребра) на завершение цикла по числу ребер; такая проверка выполняется только для отрицательных значений высот, а они могут составлять всего несколько процентов от общего числа узлов;

– однократным просмотром всех узлов для поиска всех горизонталей с заданной высотой.

Реализация всех указанных особенностей алгоритма в АСК-1 позволила увеличить его быстродействие примерно в 15 раз по сравнению с первоначальным вариантом.

Помимо прочего, использование целых чисел длиной 2 байта позволяет либо вдвое увеличить число узлов сетки при фиксированном объеме оперативной памяти, либо уменьшить необходимый объем оперативной памяти при фиксированных размерах сетки.

Дальнейшее увеличение быстродействия может быть достигнуто путем использования дополнительной оперативной памяти. С этой целью создаются два вспомогательных массива. Число элементов каждого массива равно числу узлов по оси y (числу узлов в строке). Каждому элементу обоих массивов ставится в соответствие одна строка узлов. Элементы первого массива указывают минимальное значение высоты в строке, а элементы второго – максимальное. Поиск начала замкнутой горизонтали по строке выполняется только тогда, когда она проходит между минимальным и максимальным значениями.

Упрощения алгоритма, о которых говорилось выше, заключаются в том, что не были рассмотрены случаи, когда две, три или все четыре вершины некоторого квадрата имеют значения высот, равные значению высоты отслеживаемой горизонтали. Решение в таких случаях может заключаться в том, что значения высот в узлах сетки изменяются на малую величину, которой можно пренебречь.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Алберг Дж., Нильсон Э., Уолш Дж. Теория сплайнов и ее приложения. – М.: Мир, 1972. – 316 с.
2. Аронов В.И. Методы математической обработки геологических данных на ЭВМ. – М.: Недра, 1977. – 168 с.
3. Бахвалов Н.С. Численные методы. – М.: Наука, 1975. – 631 с.
4. Баяковский Ю.М., Сорвачев А.М. ГРАФОР: комплекс графических программ на ФОРТРАНе. – М.: ИПМ, 1977. – Вып. 9. – Препринт № 102. – 41 с.
5. Белоносов А.С., Цецохо В.А. Вычислительный алгоритм и процедуры сглаживания функций, заданных приближенно в узлах нерегулярной сетки на плоскости // Некорректные задачи математической физики и проблемы интерпретации геофизических наблюдений. – Новосибирск: ВЦ СО АН СССР, 1976. – С. 6–29.
6. Бойко А.В. Методы и средства автоматизации топографических съемок. – М.: Недра, 1980. – 222 с.
7. Василенко В.А. Сплайн-функции: теория, алгоритмы, программы. – Новосибирск: Наука, 1983. –214 с.
8. Васмут А.С. Моделирование в картографии с применением ЭВМ. – М.: Недра, 1983. – 200 с.
9. Веселов В.В., Гонтов Д.П., Пустыльников Л.М. Вариационный подход к задачам интерполяции физических полей. – М.: Наука, 1983. – 120 с.
10. Вовк И.Г., Костына Ю.Г. Об аппроксимации рельефа рядом Фурье по системе ортогональных функций // Изв. вузов. Геодезия и аэрофотосъемка. – 1981. – № 4. – С. 19–25.
11. Горбик М.Д., Возгрина Л.А. Метод конечного элемента для аналитического описания топографической поверхности // Инженерная геодезия. – 1974. – Вып. 16. – С. 64–67.
12. Гэри М., Джонсон Д. Вычислительные машины и труднорешаемые задачи. – М.: Мир, 1982. – 416 с.
13. Евенко Л.И., Кочетков Г.Б. Большой бум малых ЭВМ // США – экономика, политика, идеология. – 1980. – № 1. – С. 98–107.
14. Завьялов Ю.С. Сплайн-функции – универсальный математический аппарат для представления и обработки геометрической информации в машиностроении // Вычислительные системы. – Новосибирск, 1976. – Вып. 68. – С. 3–32.
15. Завьялов Ю.С., Квасов Б.И., Мирошниченко В.Л. Методы сплайн-функций. – М.: Наука, 1980. – 352 с.
16. Иванов А.М. Исследования по математическому моделированию рельефа местности: Дисс. на соиск. учен. степ. канд. техн. наук. – Новосибирск, 1979. – 133 с.
17. Коробочкин М.И., Кронгауз А.Л. Аппроксимация топографической поверхности мультиквадриковым методом в условиях плоскоравнинного рельефа // Науч. тр. МИИЗ. – 1978. – Вып. 95. – С. 62–70.

18. Корчагин Е.К. Математическое моделирование топографических поверхностей // Изв. вузов. Геодезия и аэрофотосъемка. – 1975. – № 1. – С. 93–100.

19. Криницкий Н.А., Миронов Г.А., Фролов Г.Д. Автоматизированные информационные системы. – М.: Наука, 1982. – 381 с.

20. Малявский Б.К., Струченков В.Н. О моделировании рельефа земной поверхности поликвадратическими функциями // Изв. вузов. Геодезия и аэрофотосъемка. – 1975. – № 6. – С. 31–36.

21. Марчук Г.И. Методы вычислительной математики. – М.: Наука, 1980. – 535 с.

22. Мительман Е.Я. Исследование метода аппроксимации. – ЦНИИГАиК, 1975. – С. 76–87.

23. Моритц Г. Введение в интерполяцию и аппроксимацию. Перевод № 9107. – Новосибирск: ГПНТБ СО АН СССР, 1981. – 39 с.

24. Основные положения по созданию топографических планов масштабов 1 : 5 000, 1 : 2 000, 1 : 1 000, 1 : 500. ГКИНП – НТА-02-118. – М.: ГУГК, 1979. – 17 с.

25. Алгоритмы сплайн-аппроксимации функций в двумерных областях на системе случайно расположенных узлов в применении к обработке осредненных метеоданных над бассейном Черного моря (1946–1955): Отчет о НИР. – Новосибирск: ВЦ СО АН СССР, 1979. – 46 с.

26. Принс М.Д. Машинная графика и автоматизация проектирования. – М.: Советское радио, 1975. – 232 с.

27. Проворов К.Л., Иванов А.М. Математическое моделирование рельефа местности с использованием кубических и бикубических сплайнов // Геодезия и картография. – 1978. – № 8. – С. 39–44.

28. Справочник геодезиста (в двух книгах). – М.: Недра, 1975. – 1056 с.

29. Стечкин С.Б., Субботин Ю.Н. Сплайны в вычислительной математике. – М.: Наука, 1976. – 248 с.

30. Управление, информация, интеллект / Под ред. Берга А.И. и др. – М.: Мысль, 1976. – 383 с.

31. Хейфец Б.С. Аппроксимирование топографической поверхности ортогональными полиномами Чебышева // Изв. вузов. Геодезия и аэрофотосъемка. – 1964. – Вып. 2. – С. 78–86.

32. Цифровое моделирование местности в топографо-геодезических целях // Обзор ОНТИ ЦНИИГАиК. – 1980. – № 44. – 60 с.

33. Шульмин М.В., Мительман Е.Я. Мультиквадриковый метод аппроксимации топографической поверхности // Геодезия и картография. – 1974. – № 2. – С. 48–56.

34. Элюким С.Б., Горбушин В.П. Цифровая модель рельефа местности и ее структура // Геодезия и картография. – 1974. – № 7. – С. 36–45.

35. De Floriani L. Surface representations based on triangular grids. The Visual Computer, 3, pp. 27–50.

36. Hardy R.L. Multiquadric equations of topography and other irregular surfaces. Journal of geophysical research, 1971, vol. 76, № 8, pp. 1905 – 1915.

37. Jones C.B., Kidner D.B., Ware J.M. The implicit triangulated irregular network and multiscale spatial databases. The Computer Journal, 1994, vol. 37, № 1, pp. 43–57.

38. Schweikert D.G. An interpolation curve using a spline in tension. J. Math. Phys., 45, (1966), pp. 312–317.

39. Мусин О.Р. Структурные линии и цифровые модели рельефа // Взаимодействие картографии и геоинформатики (к 60-летию профессора С.Н. Сербенюка) / Под ред. А.М. Берлянта и О.Р. Мусина. – М.: Научный мир, 2000. – С. 21–34.

40. Кошель С.М. Цифровое моделирование и анализ геополей с помощью пакета «МАГ» // Взаимодействие картографии и геоинформатики (к 60-летию профессора С.Н. Сербенюка) / Под ред. А.М. Берлянта и О.Р. Мусина. – М.: Научный мир, 2000. – С. 41–49.

41. Скворцов А.В. Триангуляция Делоне и ее применение. – Томск: Изд-во Томского ун-та, 2002. – 128 с.

42. Васмут А.С., Гусев А.В., Кравченко Ю.А. К вопросу классификации систем цифрового моделирования рельефа местности // Геодезия и картография. – 1982. – № 2. – С. 53–57.

43. Кравченко Ю.А. Методы моделирования топографических поверхностей: Обзорная информация. – М.: ЦНИИГАиК, 1984. – 68 с.

44. Кравченко Ю.А. Определение структурных линий и точек топографических поверхностей // Сб. науч. тр. НИИПГ. – М.: ЦНИИГАиК, 1985. – Вып. 8. – С. 117–123.

45. Кравченко Ю.А. К вопросу о выборе цифровой модели высот // Вопросы картографии. Межвузовский сб. – Новосибирск, 1985. – С. 17–24.

46. Кравченко Ю.А. Моделирование топографических поверхностей с помощью интерполяционных сплайнов на подпространстве // Сб. науч. тр. НИИПГ. Автоматизация крупномасштабного картографирования. – М.: ЦНИИГАиК, 1985. – Вып. 10. – С. 42–51.

47. Кравченко Ю.А. Волновой алгоритм построения триангуляционного покрытия // Сб. науч. тр. НИИПГ. Крупномасштабные топографические съемки. – М.: ЦНИИГАиК, 1987. – Вып. 11. – С. 51–59.

48. Кравченко Ю.А. Волновые алгоритмы построения плоской триангуляции // Геодезия и картография. – 2005. – № 2. – С. 25–32.

49. Кравченко Ю.А. Компактное представление плоской триангуляции // Изв. вузов. Строительство. – 2008. – № 4. – С. 99–103.

50. Кравченко Ю.А. Оценка сложности плоских кривых и топографических поверхностей // Изв. вузов. Строительство. – 2008. – № 6. – С. 99–104.