TABLE 6.1: Sample GPS Data

	LOCATION	Garmin eTrex Personal Navigator					Garmin GPS 48				
		Longitude(DM)	atitude (DM)	Longitude (DD)	Latitude (DD)	Elev (ft.)	Longitude(DM)	atitude(DM)	Longitude (DD)	Latitude (DD)	Elev (ft.)
1	SLIP 1J33, S.B. HARBOR, CA	-119.41396	34.24263	-119.6899333	34.4043833	19.0	-119.41398	34.24264	-119.6899667	34.4044000	39.0
2	SLIP 4A48, S. B. HARBOR, CA	-119.41496	34.24389	-119.6916000	34.4064833	5.0	-119.41496	34.24389	-119.6916000	34.4064833	23.0
3	YACHT CLUB, S. B. HARBOR, CA	-119.41550	34.24209	-119.6925000	34.4034833	-7.0	-119.41551	34.24210	-119.6925167	34.4035000	47.0
4	ROB'S DRY DOCK, S. B. H., CA	-119.41493	34.24292	-119.6915500	34.4048667	-26.0	-119.41495	34.24291	-119.6915833	34.4048500	28.0
5	SLIP 3B30, S. B. HARBOR, CA	-119.41555	34.24367	-119.6925833	34.4061167	-18.0	-119.41553	34.24366	-119.6925500	34.4061000	18.0
6	S. B. MARITIME MUSEUM	-119.41622	34.24253	-119.6937000	34.4042167	-4.0	-119.41620	34.24253	-119.6936667	34.4042167	1.0
7	BREAKWATER, S. B. HRBR., CA.	-119.41289	34.24292	-119.6881500	34.4048667	23.0	-119.41286	34.24286	-119.6881000	34.4047667	36.0
8	SEA LANDING, S.B. HARBOR, CA	-119.41463	34.24478	-119.6910500	34.4079667	-16.0	-119.41461	34.24480	-119.6910167	34.4080000	12.0
9	BEACH HOUSE, S. B., CA	-119.42896	34.23774	-119.7149333	34.3962333	5.0	-119.42900	34.23773	-119.7150000	34.3962167	119.0
10	LEADBETTER BEACH, S. B., CA	-119.42149	34.23949	-119.7024833	34.3991500	6.0	-119.42148	34.23951	-119.7024667	34.3991833	45.0
11	STATE ST. FOUNTAIN, S. B., CA	-119.41338	34.24721	-119.6889667	34.4120167	27.0	-119.41338	34.24724	-119.6889667	34.4120667	50.0
12	END STEARN'S WHARF, S. B., CA	-119.41102	34.24488	-119.6850333	34.4081333	27.0	-119.41399	34.24491	-119.6899833	34.4081833	45.0
13	CAMPBELL HALL, U. C. S. B., CA	-119.50727	34.24972	-119.8454500	34.4162000	93.0	-119.50729	34.24972	-119.8454833	34.4162000	100.0
14	01 ELLISON HALL, U. C. S. B., CA	-119.50703	34.24940	-119.8450500	34.4156667	27.0	-119.50698	34.24940	-119.8449667	34.4156667	74.0
15	06 ELLISON HALL, U. C. S. B., CA	-119.50717	34.24935	-119.8452833	34.4155833	52.0	-119.50717	34.24938	-119.8452833	34.4156333	100.0
16	BIKESHOP, U. C. S. B., CA	-119.50962	34.24864	-119.8493667	34.4144000	26.0	-119.50961	34.24869	-119.8493500	34.4144833	79.0
17	BASE STORKE TWR, U. C. S. B.,	-119.50905	34.24747	-119.8484167	34.4124500	82.0	-119.50901	34.24748	-119.8483500	34.4124667	76.0
18	TOP STORKE TWR, U. C. S. B.,	-119.50895	34.24762	-119.8482500	34.4127000	-66.0	-119.50905	34.24771	-119.8484167	34.4128500	66.0
19	RINCON BEACH, CA	-119.28142	34.22571	-119.4690333	34.3761833	7.0	-119.28142	34.22571	-119.4690333	34.3761833	48.0
20	IN 'N OUT BURGER, VENTURA, CA	-119.16354	34.16099	-119.2725667	34.2683167	15.0	-119.16353	34.16101	-119.2725500	34.2683500	42.0
21	PT. MUGU, CA	-119.04804	34.05978	-119.0800667	34.0996333	42.0	-119.04805	34.05978	-119.0800833	34.0996333	97.0
22	VENTURA COUNTY LINE, CA	-118.57759	34.03161	-118.9626500	34.0526833	75.0	-118.57756	34.03165	-118.9626000	34.0527500	75.0
23	TRANCAS MKT., MALIBU, CA	-118.50567	34.01887	-118.8427833	34.0314500	15.0	-118.50573	34.01890	-118.8428833	34.0315000	54.0
24	ZUMA BEACH, MALIBU, CA	-118.48108	34.00755	-118.8018000	34.0125833	0.0	-118.49077	34.00757	-118.8179500	34.0126167	-138.0
25	LITTLE DUME BEACH	-118.47648	34.00642	-118.7941333	34.0107000	4.0	-118.47639	34.00647	-118.7939833	34.0107833	89.0
26	PARADISE COVE, MALIBU, CA	-118.47469	34.00876	-118.7911500	34.0146000	35.0	-118.47471	34.00880	-118.7911833	34.0146667	69.0
27	POINT DUME, MALIBU, CA	-118.48437	34.00185	-118.8072833	34.0030833	142.0	-118.48453	34.00197	-118.8075500	34.0032833	60.0
28	WESWARD BEACH, MALIBU, CA	-118.48624	34.00272	-118.8104000	34.0045333	50.0	-118.48623	34.00272	-118.8103833	34.0045333	99.0
29	ZUMA SUSHI REST. MALIBU, CA	-118.48819	34.01139	-118.8136500	34.0189833	118.0	-118.48816	34.01139	-118.8136000	34.0189833	114.0
						758.0					1567.0

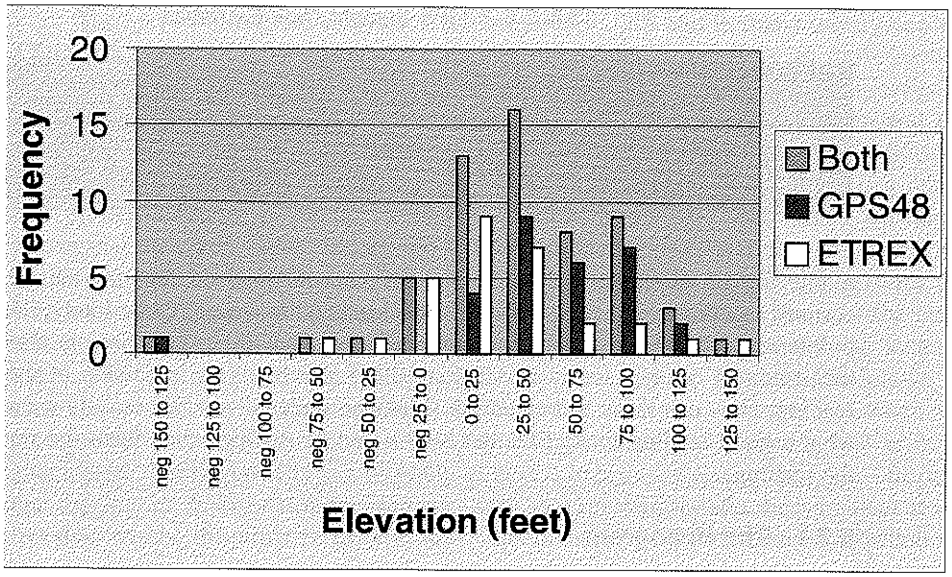


FIGURE 6.1: Histogram of the GPS elevation data from Table 6.1.

of the position as shown by the GPS receiver, and the elevation of the location. Note that the elevations are in feet, the units of data collection, and that the original GPS latitude/longitude readings in degrees-minutes format have been converted (using a spreadsheet) to decimal degrees.

6.2 STATISTICAL ANALYSIS

A first question that might be asked about the database is, What are the extremes of the data? Data extremes are simply the highest and lowest values for all records for one attribute; that is, the high and low within a single column. Table 6.2 lists these extremes and the range, defined as the highest value minus the lowest value, stated in the units of the attribute. The range of latitudes was 0.4013 degrees for the etrex, and 0.4011 for the GPS40. The range of longitudes was 0.8988 degrees for the etrex and 0.8988 for the GPS40, different only at the seventh decimal place. There was more variation in the elevation readings, which are less accurately determined with hand-held GPS units. They varied from 208 feet (63.4 m) for the etrex to 257 feet (78.3 m) for the GPS40. Both units showed elevations below sea level, highly unlikely locations for data collection, and therefore clearly errors.

We concentrate first on the elevation attribute. A first descriptive question about the data beyond the ranges is: What are the elevations of the point that were sampled? Even though most of the readings were taken along the coast, the values for the elevation range considerably. Some of the data are clearly poor readings, outside the range we would normally expect and perhaps due to the positions of the satellites or some other factor. For example, the base of the Storke tower on the UCSB campus was shown as at elevation 82 feet by the etrex and at 76 feet by the GPS48. Yet the elevation

TABLE 6.2: GPS Sample Data Extremes

Data Extremes GPS Information						
	etrex			GPS48		
	Longitude	Latitude	Elev	Longitude	Latitude	Elev
Minimum	-119.6899333	34.0030833	-66.0	-119.6899667	34.0032833	-138.0
Maximum	-118.7911500	34.4043833	142.0	-118.7911833	34.4044000	119.0
Range	0.8987833	0.4013000	208.0	0.8987834	0.4011167	257.0

of the top of the tower was −66 feet for one unit and 66 feet for the other. These curiously symmetrical readings were both clearly quite wrong! Quite likely, the reception problems of the GPS system had something to do with it, but there is also clearly a measurement error in terms of accuracy that far exceeds the precision of the elevation reading. How can these bad elevation values be screened out? Obviously we need to see what a good reading looks like and how it can be distinguished from the remainder of the readings.

6.2.1 The Histogram

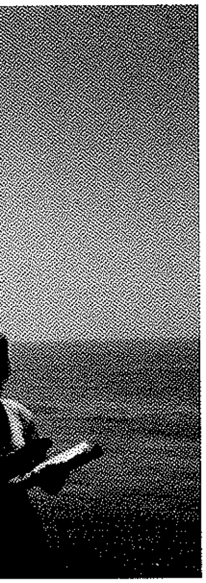
The diagram in Figure 6.1 is called a histogram. It is a plot of the data from Table 6.1 in the elevation columns. On the horizontal, bottom axis, the histogram shows the values of elevation, grouped into categories by increments of 25 feet. The actual data in the table are in feet, given to the nearest foot, which is the precision of both of the GPS receivers. On the left hand, vertical axis, we show how many elevation records fall within each elevation range. The number in each group is called that group's frequency.

FIGURE 6.2: UCSB geography student Westerly Miller collecting the data shown in Table 6.1. Point 9, Shoreline Drive, Santa Barbara, CA. (Used with Permission.)

Latitude	Elev
34.0032833	-138.0
34.4044000	119.0
0.4011167	257.0

eet for the other. These
Quite likely, the reception
ut there is also clearly a
recision of the elevation
Obviously we need to see
ed from the remainder of

he data from Table 6.1 in
gram shows the values of
e actual data in the table
oth of the GPS receivers.
records fall within each
's frequency.



a shown in Table 6.1. Point 9,

The sort of histogram shown is very common in GIS and in many other sciences. There is a cluster of values around the middle and a rapid drop-off as we move toward very high and very low values. In terms of elevation, clearly there are two sources of the variation in the readings, the real differences in elevations at the points, and the error in GPS measurements. Since none of the readings were taken underwater (although we may have been below the datum used, NAD83), clearly the error group is a major part of the variation. If we had a histogram showing only the error, then we might expect that the errors were random and would lead to a symmetrical histogram, peaking at the average or mean value. Although we took readings at only 29 points with the GPS receivers, if we had taken millions of them, it is likely that the shape of the error histogram would have become symmetrical about a central axis through an average value. This distinctive shape, called a bell curve because of its shape, is known in statistics as the normal distribution. It is the scatter of values that we get when we take measurements of a number, expecting a single value, but knowing that error is anticipated, that the error is just as likely to give a high measurement as a low one, and that the amount of error is not systematic, such as a misreading of a measurement scale. In other words, the error is random and unpredictable. Many real-world distributions show this form.

Other types of error distributions are shown in Figure 6.3. The histogram could be skewed to one side, which would mean that it is more likely to over- or underestimate a number than the opposite. The values could be equally even and dispersed about the average, implying that the measurement is perfectly accurate or at least perfectly consistently wrong. The error could be the same everywhere along the line, implying that no value is a better reading than any other. We could get a group of errors, perhaps occasional misreadings of a 3 as a 5 and the occasional dropping of a decimal place. This would give us a multi-peaked histogram, as shown in Figure 6.3.

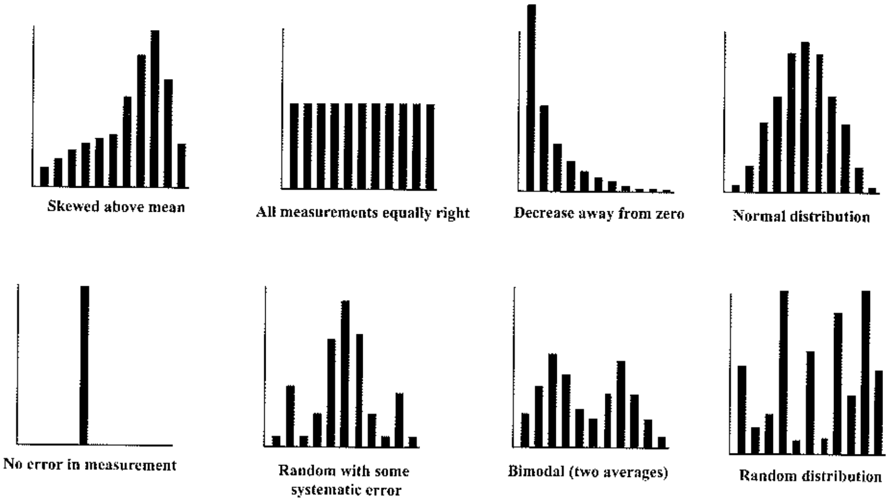

Skewed above mean
All measurements equally right
Decrease away from zero
Normal distribution
No error in measurement
Random with some systematic error
Bimodal (two averages)
Random distribution

FIGURE 6.3: Some possible alternatives to the normal distribution with a large number of observations.

6.2.2 The Mean

If errors are consistent, we can correct for them. For example, we could reason that all the negative elevations in Table 6.1 were in error. We could safely eliminate these seven numbers and probably correctly believe them to be wrong. We do not have this option if we have only one number or reading—we have no choice but to use it! If we have two readings only, and they disagree, we would probably average them. If we had three readings and they disagreed, we could average them, reject one reading that was obviously wrong (varies by too much), or average the two readings that most closely agree with each other.

The more numbers we have, the more we can see what the typical amount of variation is, that is, how corrupted are the readings by a random amount of error. If this is the case, as in Figure 6.1, then we can go ahead and average the numbers, or at least give expected amounts of error. There are alternatives to averaging. For example, a simple representative value for a group of records can be selected by sorting the elevations by height, which the GIS database manager can do (Table 6.3), and then taking the value of the middle reading. This value is called the median. This works fine for an odd number of elevation readings; the center comes out exactly. An advantage is that this is a "real" typical value because it is an actual part of our data set. If the attribute in the database was, for example, state average salary for GIS professionals in dollars, we could pick the middle or median state and compare our own state to it. This is not so simple if we have an even number of records, in this case 50. We have to take the two center values and average them, losing the attachment to a single data record.

In the case of the GPS data we have a total of 58 elevations, an even number. Table 6.3 sorts these values, and highlights the two at the center. These are 39 and 42 feet, which average to (39 + 42) ÷ 2 or 40.5 feet. The median elevation of the data in Table 6.1, therefore, is 40.5 feet. The median is called a measure of central tendency, because it gives us a value which is descriptive of the group of values as a whole. A nice feature of the median is that it is unaffected by the extreme values at the limits of the distribution. Another is that it is made up of values that actually occur in the data (39 and 42), and is not an abstract number.

Another measure of central tendency is the average, known in statistics as the mean. It is computed rather simply, by adding together the values for all the records and dividing by the number of records. We may have to leave some of the numbers out of the averaging process. For example, we may choose to exclude the negative elevations from our calculations. These values will then be called missing values. It is important not to confuse missing values with zero, which is not a missing value, but a legitimate reading of "0.0."

The sum of the elevations for Table 6.1, shown at the bottom of the elevation column in Table 6.3, is 2325 feet. This gives us an average or mean elevation of 40.086 feet. Excluding the negative elevations gives a higher sum of 2645 feet for a higher average of 51.863 feet (note that the median was 40.5 feet). Because the mean was calculated by dividing, it is more precise than the readings themselves; that is, it has more significant digits. The elevations were given by the GPS receiver to the nearest foot, and we have rounded the mean to the thousandth or third decimal place. We could also say that it is more accurate, because the measurement now reflects many readings, not just one. The more readings we take at each location (we took two), the more accurate the location it should become, unless there is some systematic error of which we are unaware (say, the

e, we could reason that
ld safely eliminate these
ng. We do not have this
hoice but to use it! If we
average them. If we had
ect one reading that was
adings that most closely

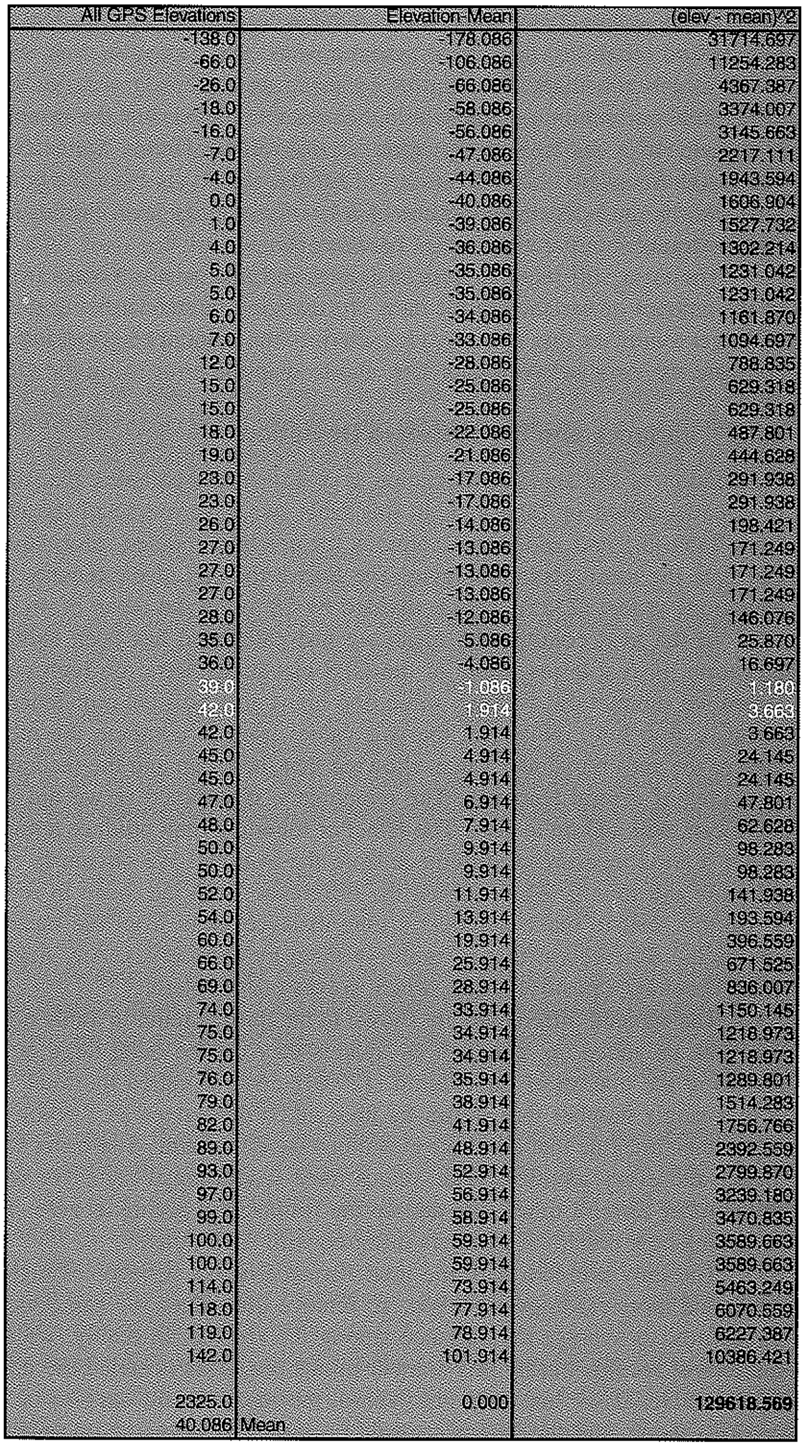
it the typical amount of
n amount of error. If this
e the numbers, or at least
ng. For example, a simple
sorting the elevations by
l then taking the value of
s fine for an odd number
ige is that this is a "real"
attribute in the database
n dollars, we could pick
his is not so simple if we
ike the two center values
l.

vations, an even number.
enter. These are 39 and
lian elevation of the data
asure of central tendency,
of values as a whole. A
ne values at the limits of
ctually occur in the data

nown in statistics as the
es for all the records and
ne of the numbers out of
le the negative elevations
ng values. It is important
ng value, but a legitimate

ttom of the elevation col-
n elevation of 40.086 feet.
feet for a higher average
e mean was calculated by
is, it has more significant
nearest foot, and we have
e could also say that it is
eadings, not just one. The
re accurate the location it
we are unaware (say, the

TABLE 6.3: Variance in the Elevation Data



All GPS Elevations	Elevation-Mean	(elev - mean)^2
-138.0	-178.086	31714.697
-66.0	-106.086	11254.283
-26.0	-66.086	4367.387
-18.0	-58.086	3374.007
-16.0	-56.086	3145.663
-7.0	-47.086	2217.111
-4.0	-44.086	1943.594
0.0	-40.086	1606.904
1.0	-39.086	1527.732
4.0	-36.086	1302.214
5.0	-35.086	1231.042
5.0	-35.086	1231.042
6.0	-34.086	1161.870
7.0	-33.086	1094.697
12.0	-28.086	788.835
15.0	-25.086	629.318
15.0	-25.086	629.318
18.0	-22.086	487.801
19.0	-21.086	444.628
23.0	-17.086	291.938
23.0	-17.086	291.938
26.0	-14.086	198.421
27.0	-13.086	171.249
27.0	-13.086	171.249
27.0	-13.086	171.249
28.0	-12.086	146.076
35.0	-5.086	25.870
36.0	-4.086	16.697
39.0	-1.086	1.180
42.0	1.914	3.663
42.0	1.914	3.663
45.0	4.914	24.145
45.0	4.914	24.145
47.0	6.914	47.801
48.0	7.914	62.628
50.0	9.914	98.283
50.0	9.914	98.283
52.0	11.914	141.938
54.0	13.914	193.594
60.0	19.914	396.559
66.0	25.914	671.525
69.0	28.914	836.007
74.0	33.914	1150.145
75.0	34.914	1218.973
75.0	34.914	1218.973
76.0	35.914	1289.801
79.0	38.914	1514.283
82.0	41.914	1756.766
89.0	48.914	2392.559
93.0	52.914	2799.870
97.0	56.914	3239.180
99.0	58.914	3470.835
100.0	59.914	3589.663
100.0	59.914	3589.663
114.0	73.914	5463.249
118.0	77.914	6070.559
119.0	78.914	6227.387
142.0	101.914	10386.421
2325.0	0.000	**129618.569**
40.086	Mean	

elevations were on the wrong datum, or the earth's fit to the ellipsoid was pretty poor at this place!).

To be accurate, we must test the value against an independent source of higher fidelity and reliability than these measurements. On a map, such an elevation could be found at a bench mark elevation or by using an extremely accurate GPS system with differential correction to achieve a higher accuracy level. For the sample data, we could plot them on a digital raster graphic version of the United States Geological Survey quadrangle maps at 1 : 24,000, and read off the elevations from the contours. This may, however, be an independent source of a lower authority level because the contour map would require interpolating values between contours.

The error still requires understanding, and this usually comes from a computation of the amount of variance in the data. Normalizing or standardizing this variance allows us to get a number that not only characterizes the amount of error, but also allows us to compare the error between data sets, and to map the error using the GIS.

6.2.3 Variance and the Standard Deviation

Calculation of the standard deviation is a little more complex than the mean but can be seen in action in Table 6.3. In this table, we have extracted the elevation columns from Table 6.1 and added two columns for newly computed information. First, we can calculate the mean elevation again. The line at the end of the records is the total of the elevations in feet (2325). This is divided by 58 to get the mean of 40.086 feet.

In the next column, the mean we just calculated has been subtracted from each elevation. In real terms, some of the GPS elevations were too high and some too low. When we subtract the average, some (too high) give positive remainders, and some give negative (too low). Note that the two values used to compute the median were either side of the zero remainder value, showing that the mean and the median are indeed both measures of central tendency. The remainders, if summed as we did for calculation of the mean, will add up to zero. The effect of rounding may produce a nonzero result. This is not a good measure of difference from the average! We obviously need to get rid of the plus and minus. We need just the difference from the mean in feet. One easy way to do this, and to make numbers a long distance away from the mean stand out, is to square the differences, that is, multiply them by themselves. Any negative value times itself becomes a positive value, which solves our problem.

The squared differences are shown in the third column in Table 6.3. These numbers can now be added up, although as squares they can be quite large numbers. The sum, shown at the bottom of the column (129,618.569), is called the total variance, because it is a single number that shows by how much the values as a whole disagree. If you doubt this, imagine that all of the elevations were the same. The mean would be the same value, and every identical value minus itself would equal zero. We would then square the number zero 58 times, and add up the results, still giving us zero. Thus the variance measures how much the numbers disagree with each other.

The only problem with the total variance is that it becomes higher and higher as more and more readings are taken. To remove this effect, we can divide by the number of records, to average the variance among the records. In reality, we divide not by the number of records but by this number minus one. This is called the degrees of freedom of the value. If we had only one record, the mean would have to be the same as this record's value, and the variance would be zero. We get variation only when the second value is measured, so we say that with two readings we have one degree of freedom. Dividing 129,618.569 by 57 (58 − 1) gives a value of approximately 2274.01.

In addition, we can take the square root of the result so that the units are back as feet rather than as square feet. This number (47.687), standardized in this way, is called the standard deviation. This number is in the same units as the values of the records, in this case feet. Better than just that, it is the average amount by which readings differ from the average and so can be called an expected error of any given reading. Of course, the readings are just as likely to be above the average as below. We could say that the elevation within the point sample is the mean (40.1 feet) plus or minus the expected error of 47.7 feet, and so is most likely to lie between −7.6 feet and 87.8 feet. These limits are called the error band or margin of error.

6.2.4 Statistical Testing

The final step that we can put these elevation values through is to do statistical tests. Important for this stage is the idea that the measurements we took with the GPS receiver are just a few of a large number of possible readings. In statistical jargon, the entire possible set of GPS readings is called a population, and the actual numbers in Table 6.1 are a sample from this population. For most statistical purposes, the sample is considered as only a tiny proportion of the population as a whole. For example, we could consider our sample as representative of all coastal area elevations in Southern California.

The purpose of drawing this division is that we can use the sample's mean and variance to make an estimate of those of the population, and then use the two interchangeably. The reason we divided by 57 and not 58 to get the variance is because we want a population, not a sample estimate. It is useful to us, for example, to know the standard deviation of the elevation values at the location given in Table 6.1, because we can use it to get an estimate of the difference in elevation between the two GPS receivers compared with the overall measurement capability of GPS as a way to record elevation in the sample.

We can use a statistical model of the bell curve, called the standard normal distribution, to estimate how likely any given measurement of the elevation is to be correct. This distribution, provided in most statistical textbooks, allows us to look up the standard deviation and the number of records and to estimate the odds against getting this elevation measurement given the tabulated standard deviation. The numbers in the statistical table are the amount of area beneath the standard normal curve that corresponds to probabilities. This is one way that we can determine whether the largest and smallest measured elevations are reasonable.

For example, if our mean elevation is 40.01 feet and the standard deviation is 47.2 feet, what is the chance of getting a GPS reading of −138 feet? The reading of negative 138 feet is 178.1 feet below the mean, or 178.1 ÷ 47.2 of a standard deviation. This number of standard deviations is called a Z-score. In this case it is 3.77. Looking this up in the statistical table shows us that 0.4999 of the curve lies between the mean and this value, and 0.0001 lies beyond it. The normal curve is symmetrical about the mean. These values are shown in Figure 6.3.

The standard normal distribution table tells us that for our 58 readings, the likelihood of a number falling at random between the mean and this value is 49.99%. This is obviously large compared to the 0.01% chance of falling lower. So we could argue that there is a strong probability that this reading is in error. Often, a reading is rejected as highly unlikely if it is more than two standard deviations away from the mean.

Another way that these numbers can help in analysis is to answer the following question: Is there a difference in measured elevation between the two makes of GPS receiver? It would be quite simple to calculate the mean separately for each of the two types of receiver. In the GIS database we could select out those attributes based on their "GPS-type" attribute. Similarly, we can also recalculate the standard deviations for the two data sets separately. When we do this, we get for the Garmin eTrex a mean of 26.14 feet and a standard deviation of 43.71 feet, and for the Trimble GPS, a mean of 54.03 feet and a standard deviation of 48.12 feet.

Now we can return to the question of whether the two values differ from each other. If they were the same, they would have the same means exactly. However, because the two sample means were estimates of the real elevations using 58 points, we have to compare the difference between the two means against the estimated population standard deviation. This is called a test of means. As the number of samples is different, we have to estimate the standard deviation of the overall mean as the square root of the two normalized variances added together. For the eTrex, the sum of squared differences from the mean was 53,497.36, and dividing by the 28 (29 minus 1) samples gives us a normalized variance of 1910.62. Similarly, for the GPS48, the sum of squared differences from the mean was 64,836.8, and dividing by the 28 (29 minus 1) samples gives us a normalized variance of 2315.6. Adding these numbers and taking the square root of the results gives almost exactly 65 feet.

This number is the estimated standard deviation of the mean itself. This implies that we can test the difference between the means against it, computing a Z-score and a probability as before. The two means of 26.14 feet and 54.03 feet differ by 27.89 feet. Dividing by the calculated standard deviation of the mean gives a number of 0.429. This value cannot be looked up in the regular statistical table but can be compared on a scale called Student's T. We also have to look up the T value by its number of degrees of freedom, which here is the total of the two samples minus two, or 50. This is another probability table. At the 5% confidence level, T would have to be about 1.675 or greater to reject the hypothesis that the two means were the same.

At only 0.429, then, we cannot prove statistically that there is any difference between the elevations measured by the two GPS receivers. Thus, even though eTrex has a lower standard deviation, and the GPS48 seems to often give readings that appear high for a coastal area, neither gives elevation values superior to those of the other. The conclusion, then, is that the two GPS receivers are similar in capability, with an expected error in estimating elevations of their "population" average of 65 feet. This is hardly what could be described as cartographic precision. The results may also reflect the influence of a few large errors or blunders. Fortunately, the receivers are better at latitude and longitude.

The standard deviations are also similar, but differ by 4.41 feet. These can be thought of as the accuracy of the elevation for the two GPS receivers. The slightly better (lower) value for the eTrex may be due to several things: the technical characteristics of the receiver, the fact that the receivers have different precision, even perhaps an interaction between the two receivers. Statistics can also allow us to test whether the two standard deviations are significantly different from each other, or whether they could be different because of random causes. This, however, we leave for later, or for a

ı, a reading is rejected as
/ from the mean.
to answer the following
ı the two makes of GPS
rately for each of the two
e attributes based on their
andard deviations for the
Garmin eTrex a mean of
Trimble GPS, a mean of

ues differ from each other.
tly. However, because the
ıg 58 points, we have to
mated population standard
samples is different, we
as the square root of the
um of squared differences
nus 1) samples gives us a
sum of squared differences
nus 1) samples gives us a
king the square root of the

mean itself. This implies
computing a Z-score and a
3 feet differ by 27.89 feet.
es a number of 0.429. This
an be compared on a scale
its number of degrees of
wo, or 50. This is another
o be about 1.675 or greater

at there is any difference
Thus, even though eTrex
ı give readings that appear
rior to those of the other.
ilar in capability, with an
average of 65 feet. This is
he results may also reflect
the receivers are better at

y 4.41 feet. These can be
PS receivers. The slightly
gs: the technical character-
nt precision, even perhaps
o allow us to test whether
ach other, or whether they
we leave for later, or for a

statistics class, because we have yet to look at the geographic, mappable characteristics of these numbers.

6.3 SPATIAL DESCRIPTION

In the preceding section we looked at how to describe a single attribute statistically. The first and most significant factor in dealing with spatial data is that there are at least two spatial measurements, an easting and a northing. We could summarize spatial description, as describing two attributes simultaneously.

In the simplest and most basic way, we can duplicate the attribute descriptions above for the locational data to give spatial descriptions. In this case, we can treat the two separate parts of the coordinates, the eastings and the northings, as if they are each a single attribute, which indeed they are. Just as we began the discussion of describing the values of a single attribute by discussing the concept of a minimum and a maximum value for an attribute and the concept of a range, when the attributes describe coordinates, a first point is described by the minimum easting and the minimum northing, and a second point describes the corresponding maxima. The two points define a rectangle, whose two side lengths are the ranges in easting and northing, respectively, and that encloses all the points.

This is called the bounding rectangle of the points. It can be found by simply sorting the records by easting, and taking the first and last record, and then repeating for the northing. A bounding rectangle for the points in Table 6.1 is included in Figure 6.4.

6.3.1 The Mean Center

In much the same way that we calculated means and standard deviations separately for the two GPS receiver's elevations, so also were they calculated for the latitudes and longitudes. These were first translated into decimal degrees, then summed and divided to find the average latitude and longitude for the eTrex (34.2840575°N, 119.4430961°W)

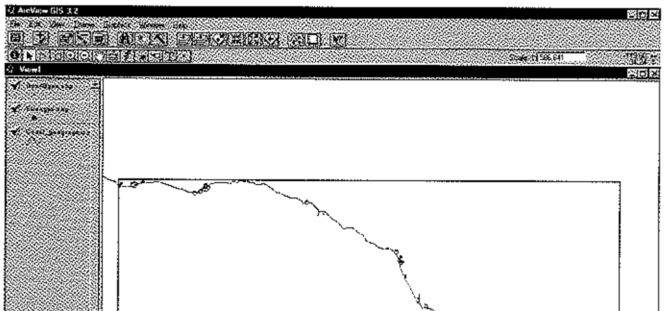


FIGURE 6.4: Bounding rectangle of the GPS data, showing the spatial extremes of the data set.

and for the GPS48 (34.2840879°N, 119.4438299°W). The result of the two means is itself a point, with both a real geographic location and a special geographic name, the mean center. This point is also sometimes called a centroid, a point chosen (in this case statistically) to represent a geographic distribution. Although the GPS data are a set of points, lines and area features can also have a centroid, selected in any one of several ways.

Figure 6.5 is a photograph of a place in Rugby, North Dakota, that claims to be the geographic center of North America. Although this is a fascinating monument and the nearby diner is probably heavily dependent on its visitors for its food business, it should be quite obvious that, unlike for a set of points, an entire continent could have any number of centroids! For example, this could be the point farthest from any coastline, the center of all of the points making up the coastline, the center of the bounding rectangle, or the center of the largest circle that can be drawn inside North America.

The World Almanac lists the geographic center of North America not in Rugby, but in Pierce County, North Dakota, 10 km west of Balta (48°10′ N, 100°10′ W). The mean

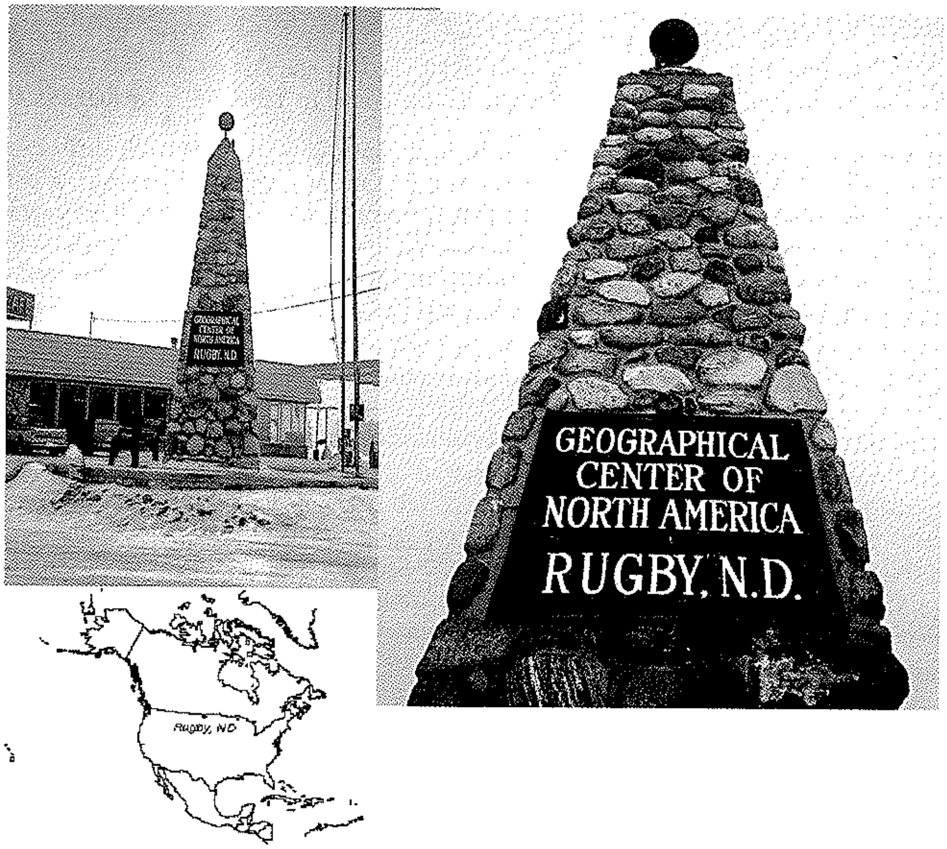
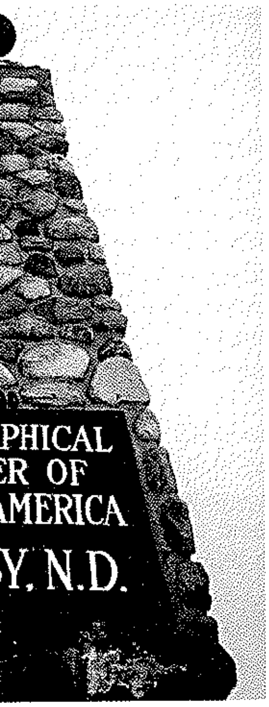


FIGURE 6.5: The monument in Rugby, North Dakota. Photographs by Colette Flanagan. (Used with permission.)

ult of the two means is
al geographic name, the
a point chosen (in this
ugh the GPS data are a
, selected in any one of

Dakota, that claims to
a fascinating monument
ors for its food business,
n entire continent could
point farthest from any
stline, the center of the
n be drawn inside North

merica not in Rugby, but
T, 100°10′ W). The mean



Colette Flanagan. (Used with

center calculation would also change depending on the map projection, datum, and ellipsoid. Judging from the flags seen on a visit to the site, it is not even clear whether Rugby's definition of North America includes Mexico, Alaska, Greenland, or Hawaii. Obviously, one place is as good as any for this type of monument. It would be interesting to know whether the diner predates the monument, or vice versa.

6.3.2 The Standard Distance

Figure 6.6 is a plot of the GPS points in map space, with a different colored symbol for each of the two GPS receivers. Looking at this map, is it possible to see any difference between the two sets of measurements? Since the overall spread of the points exceeds the differences in location between the two receivers, it is very hard to say, even on the zoom of Santa Barbara harbor. Instead, we can compare the distributions statistically, by examining the standard deviations in the easting and northing directions, in this case in latitude and longitude.

Imagine the line between the two GPS locations for each point, with all eTrex points drawn on top of each other. We could look at the bearing of these lines, "rays" stretching out between the two readings for each point. We would expect the bearings and the lengths to be random, but the average length would now give a mean with a real meaning, the expected average distance difference between the two receivers. This can also be calculated from the standard deviations in the easting and northings, calculated as the square root of the sum of the squared distances in the two directions.

Converting to a single number like this is termed normalizing. This parameter, called the standard distance, is a map equivalent of the standard deviation. Again, are there major differences between the two GPS receivers? The standard distance is at last truly a measure of the agreement between the GPS receivers. In degrees, the standard deviation in longitude was 0.003178 and in latitude was zero. Thus the two receivers differed only in longitude, by an amount that converts to a surprisingly high 352.5 meters or 1156 feet. What could have caused this difference? Calculating for each receiver separately, and using standard length tables for the length of a degree at different earth locations, the mean centers are off by 81.4 meters in their easting but only 2.8 meters in

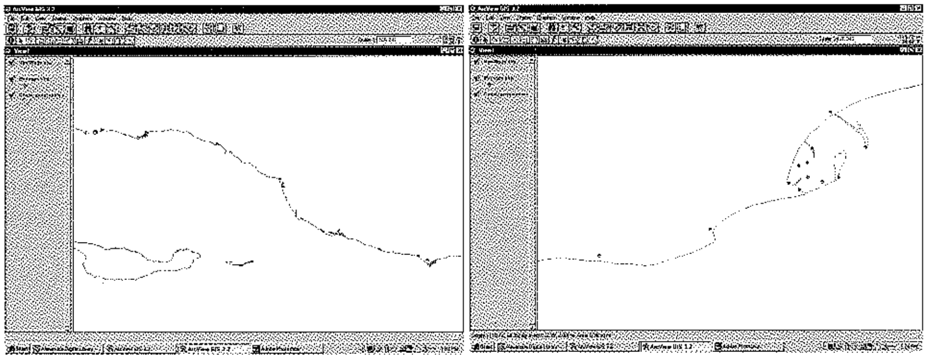


FIGURE 6.6: ArcView display of the data in Table 6.1. Both the GPS48 and the eTrex Garmin GPS receiver data are plotted, but are so close spatially that they are not separable at this scale. Area displayed is Santa Barbara to Malibu, California. Right is a zoom of Santa Barbara harbor area.

their northing. What has caused the difference: chance, systematic error, or some other cause? Just by asking this question and having some data and a method to answer it objectively, we are already doing spatial analysis.

While statistics are useful in demonstrating that an error is present, and that it has an impact on the aggregate statistical descriptors, the GIS can help us to isolate exactly which readings have caused the problem. Figure 6.7 shows a map created in ESRI's ArcView 3.2 made by bringing data from the spreadsheet created to store the GPS data as a table and mapping it. The value plotted for each point is the difference in latitude and longitude for each point, squared to get rid of negatives, divided by the size of the sample, and added for latitude and longitude. The values were converted to meters using the same tables as above, and the square root of the sum taken. This is then a map of the magnitude of the total spatial discrepancy between the two GPS receivers. Two points clearly stand out, point numbers 12 and 24. These two points alone account for almost all of the error in the data. Without them, the two receivers seem to be not only quite accurate, but also in agreement. As implied in the discussion, the largest errors are in longitude. What caused them? It seems unlikely that the errors were caused by the random errors inherent in the GPS receivers, which are of a far lesser magnitude. Could numbers have been inverted (say a 345 became a 543), digits transliterated (a "5" read as an "8"), or perhaps a digit was left out when the values were entered into the spreadsheet?

Overall, it is clear that the statement at the start of the next chapter is correct, that a map is a set of errors that have been agreed upon! On the other hand, the mapping

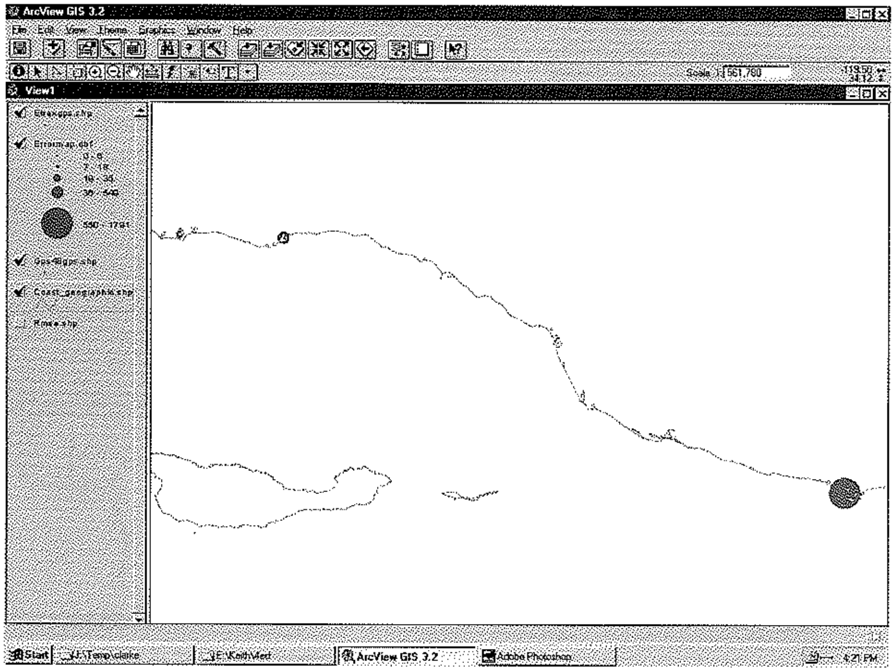
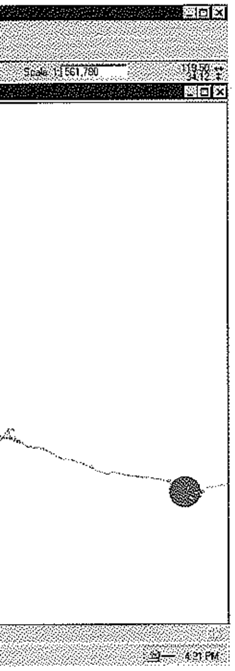


FIGURE 6.7: ArcView map of the GPS data with root mean squared error between the two GPS devices, units rescaled into meters.

tic error, or some other a method to answer it

s present, and that it has elp us to isolate exactly map created in ESRI's ed to store the GPS data he difference in latitude vided by the size of the onverted to meters using en. This is then a map wo GPS receivers. Two points alone account for ers seem to be not only ssion, the largest errors errors were caused by a far lesser magnitude. igits transliterated (a "5" es were entered into the

t chapter is correct, that other hand, the mapping



etween the two GPS devices,

capability of the GIS and the power of statistical and spatial analysis are such that we can use our same GIS tools to find errors in our data. However "pure" we believe our data to be, there are always errors. It is often said that eliminating the last 1% of error costs as much as eliminating the other 99%. It is best, therefore, to be both aware of errors and to be able to describe them as far as your own data are concerned, especially spatially.

6.3.3 Geographic Features and Statistics

In Chapter 2 we met the idea that geographic features can be classified into points, lines, and areas by their dimensions on the map. Describing each of these can lead to measuring spatial properties directly from the digital files containing the geocoded representations of the features. We started Chapter 6 with a set of points, the GPS example, because points are the easiest type of feature to describe. Although we have so far used quantitative measures to describe geographical features, many arrangements of features are described verbally.

For example, points are sparse, uneven, random, regular, uniform, scattered, clustered, shotgun, or dispersed. Patterns are regular, patchwork, repetitive, or swirling. Shapes are rounded, oval, oblong, drawn-out, or resemble Swiss cheese. The challenge is to find numbers that say the same thing. The bounding rectangle, the mean center, and measures such as the standard distance can provide excellent descriptors of points, although more complex measures are obviously needed for the higher-dimension features.

Lines have a number of points, a line length, a distance between the start and end points (or nodes), the average length of one of the line segments, and a line direction. A useful description of a line could be the ratio of actual line length divided by the start-to-end node length, what could be called a straightness index. For a straight line, this measure would be 1. For the Mississippi River this would be a far larger number. The direction could be taken as the clockwise angle bearing from north (the overall "trend" of the line), although this value would have a big variance along a curved or wiggly line, too.

Areas are even more difficult to describe. Simplest to measure with a GIS are the area in square meters, the length around the boundary, the number of points in the boundary, the number of holes, and the elongation, taken as the length of the longest line axis divided by the axis at 90 degrees to it. We can also divide the area of the bounding rectangle into the area, a space-filling index with a maximum value of 1. If the area has neighboring areas, we could count them or determine the average length of the area's boundary shared in common with a neighbor.

Not all of these numbers are easy to compute with a GIS. Sometimes a multiple-step process must be used and information created or computed in the attribute database and passed back to the map for display. Almost every GIS has a compute command that allows the performing of math operations like

```
COMPUTE ATTR5 = (ATTR2 + ATTR3)/ATTR4
```

Each new measure, however, can be an intermediate step in the computation of another statistic. For example, we could measure the lengths of streams by district in kilometers, measure the area of the district in square kilometers, and then create a new attribute in the database of the stream density in meters of stream per square kilometer. This might be an interesting data value to then map by district in the GIS. Computations are so common using areas of polygons that many GIS packages compute and save areas when they are created, whether they are wanted or not.

6.4 SPATIAL ANALYSIS

Numbers that describe features are useful, but as we noted in Chapter 2 the purpose of geographic inquiry is to examine the relationships between geographic features collectively and to use the relationships to describe the real-world phenomena that the map features represent. The geographic properties we noted in Figure 2.15 were size, distribution, pattern, contiguity, neighborhood, shape, scale, and orientation.

Each spatial relation begs three fundamental questions: (1) How can two maps be compared with each other?; (2) How can variations in geographic properties over a single area or GIS data set be described and analyzed?; and (3) How can we use what we have learned using the analysis to explain and therefore predict future maps of the geography in question? The third question may be as simple as selecting the best route from A to B on a map, or as complex as modeling the future growth of cities based on their size, shape, and development over time. GIS gives us the capability of doing both, and anything in between. In terms of comparing maps, a simple way is to bring multiple maps into coregistration and then merge their themes to make a composite. This is what is meant by map overlay analysis. An example of map overlay will follow a first discussion of spatial models and how GIS adds to their construction, examination, and use.

6.4.1 U.S. Gender Ratios: An Example

A full set of descriptive statistics for all of the properties listed here is beyond the scope of this book. Instead, two geographic analysis problems will be covered, starting with a simple geographic distribution, and ending with some speculations about prediction.

Consider the data in Table 6.4 and the accompanying maps shown in Figure 6.8. The value shown is the gender ratio, defined as the number of males in the population per 100 females. States with numbers over 100 have more men than women, and those states lower than 100 have more women than men. The numbers are selected from the 2000 census of population for the United States. After selecting the data, an obvious first step is to map the distribution. This map, using a method called *choropleth mapping*, in which states are shaded by gender ratio in groups of values, is shown as Figure 6.8 on the upper left.

Remember that looking at the map, using the full power of the human visual system and just plain thought, is every bit as powerful as any spatial analysis method. This critical step, *plot* and *look*, should never be skipped in GIS analysis, at the risk of using complex methods to prove the geographically obvious. So, first, examine the upper left map in Figure 6.8 carefully and then move on to the next section. How can this distribution be described? The data are ratio values for areas. Each geographic property suggests a question about the data, with a few of these listed below.

SIZE What are the high and low values of the gender ratio? Do large states have characteristic values? What about small states?

DISTRIBUTION Is there a regional difference between numbers? Does the West have all the low values and the East the highs? Are there clusters of similar values?

compute and save areas

Chapter 2 the purpose of
ographic features collec-
henomena that the map
re 2.15 were size, distri-
ntation.
(1) How can two maps
raphic properties over a
) How can we use what
edict future maps of the
selecting the best route
growth of cities based
the capability of doing
simple way is to bring
make a composite. This
p overlay will follow a
nstruction, examination,

here is beyond the scope
covered, starting with a
ons about prediction.
ps shown in Figure 6.8.
males in the population
than women, and those
rs are selected from the
the data, an obvious first
choropleth mapping, in
shown as Figure 6.8 on

er of the human visual
spatial analysis method.
S analysis, at the risk of
first, examine the upper
t section. How can this
ach geographic property
elow.

io? Do large states have

numbers? Does the West
lusters of similar values?

TABLE 6.4: Ratio of Males to Females Times 100 for the Lower 48 United States. Source 2000 Census

State	Males/100 Females	Longitude	Model	Residual
AL	89.00	-86.2833	95.6930	-2.3930
AR	92.50	-92.2667	96.5535	-1.2535
AZ	93.00	-112.0000	99.3913	0.3087
CA	93.10	-121.5000	100.7575	-1.4575
CO	93.30	-104.9833	98.3822	3.0178
CT	93.40	-72.6667	93.7348	0.1652
DC	93.40	-77.0000	94.3580	-5.3580
DE	93.40	-75.5000	94.1423	0.2577
FL	93.80	-84.2833	95.4054	-0.1054
GA	93.90	-84.3833	95.4198	1.3802
IA	94.30	-93.6167	96.7476	-0.4476
ID	94.40	-116.2000	99.9953	0.5047
IL	94.40	-89.6167	96.1724	-0.2724
IN	94.50	-86.1333	95.6714	0.6286
KS	94.60	-95.6833	97.0448	0.6552
KY	94.60	-84.9167	95.4965	0.1035
LA	94.80	-91.1667	96.3953	-2.5953
MA	94.90	-71.1167	93.5119	-0.5119
MD	95.30	-76.4167	94.2741	-0.8741
ME	95.30	-69.7000	93.3082	1.4918
MI	95.60	-84.5833	95.4485	0.7515
MN	95.90	-93.0833	96.6709	1.4291
MO	96.00	-92.1667	96.5391	-1.9391
MS	96.10	-90.1667	96.2515	-2.8515
MT	96.20	-112.0000	99.3913	-0.0913
NC	96.30	-78.6500	94.5953	1.4047
ND	96.30	-100.7667	97.7759	1.8241
NE	96.30	-96.7167	97.1934	0.0066
NH	96.60	-71.5000	93.5670	3.2330
NJ	96.70	-74.7667	94.0368	0.2632
NM	96.80	-106.0000	98.5284	-1.8285
NV	96.80	-119.7500	100.5058	3.3942
NY	97.20	-73.8333	93.9026	-0.8026
OH	97.60	-83.0000	95.2208	-0.8208
OK	97.70	-97.5333	97.3109	-0.7109
OR	98.10	-123.0500	100.9804	-2.5804
PA	98.40	-76.8333	94.3340	-0.9340
RI	98.50	-71.3833	93.5502	-1.0502
SC	98.60	-81.0000	94.9332	-0.4332
SD	99.10	-100.3333	97.7135	0.7865
TN	99.30	-86.8000	95.7673	-0.8673
TX	99.30	-97.7000	97.3348	1.2652
UT	99.60	-111.8667	99.3721	1.0279
VA	99.70	-77.5000	94.4299	1.8701
VT	100.40	-72.5833	93.7228	2.3772
WA	100.50	-122.8667	100.9540	-1.8540
WI	101.20	-89.3833	96.1388	1.4612
WV	101.40	-81.5833	95.0171	-0.4171
WY	103.90	-104.8167	98.3583	2.8417

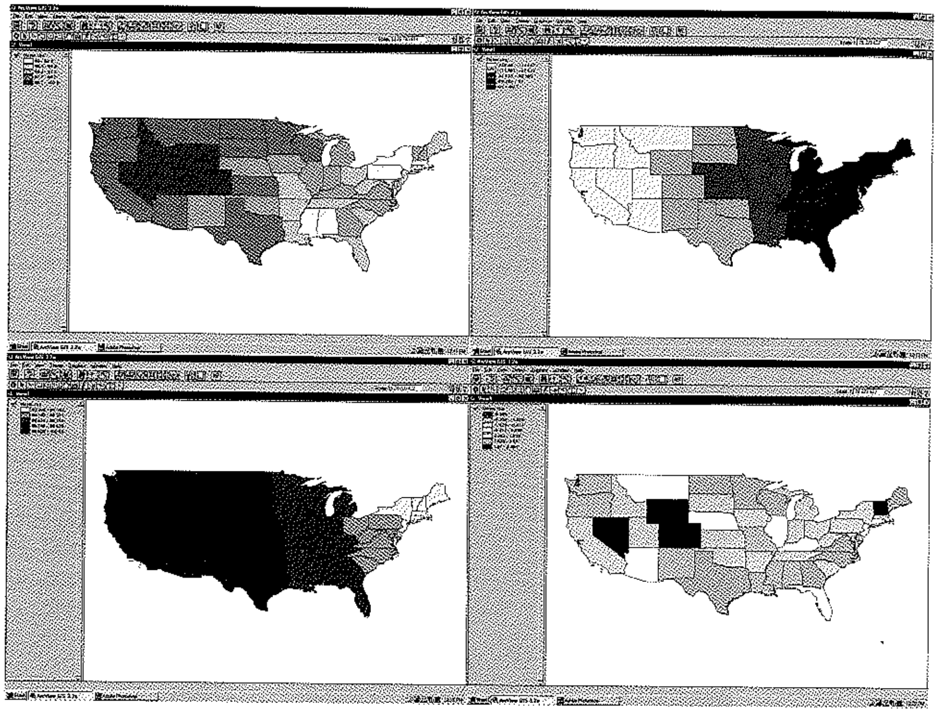


FIGURE 6.8: Data for the Gender Ratios example plotted on the USA lower-48 states base map. Upper left: the gender ratio data (2000 Census, Male to Female ratio). Upper right: Longitude of state capitals. Lower Left: The linear model. Lower Right: residuals from the model.

PATTERN Is there a distribution that repeats over the space? Are states with high gender ratios always at some regular distance from those with low gender ratios?

CONTIGUITY Are states with high gender ratios always surrounded by states with low values, or vice versa?

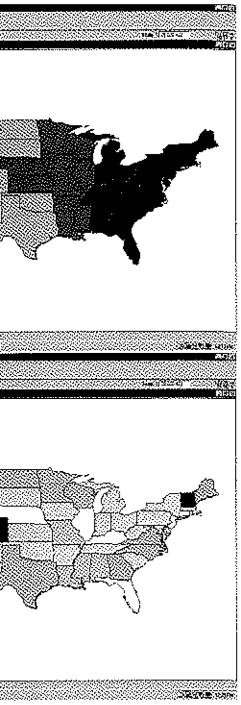
NEIGHBORHOOD Does the gender ratio drop off steeply around several key focal points?

SHAPE Do elongated states have higher gender ratios than those of rectangular or irregular states?

SCALE Would all of the other geographic properties change if we examined the gender ratio using another, higher-resolution, set of districts, such as counties or minor civil divisions?

ORIENTATION Are there connecting lines, perhaps following major highways, between states with high gender ratios going coast to coast? Is there a direction to a general increase in the gender ratio across the map?

Clearly, a whole host of statistics could be used to answer any one of these problems, and many different analyses could be performed. Not all of them will be possible with the GIS that we have, however. We will simply take the last of the questions and learn by example. First, we could divide the map into an eastern and a western half of



3 states base map. Upper left:
tude of state capitals. Lower

ce? Are states with high
vith low gender ratios?
ys surrounded by states

ply around several key

ian those of rectangular

nge if we examined the
cts, such as counties or

owing major highways,
? Is there a direction to

any one of these prob-
of them will be possible
ast of the questions and
n and a western half of

the country using the traditional dividing line of the Mississippi River. If we do this, the GIS attribute database could be edited to include a new attribute, COAST, with attribute values EAST or WEST. We could then compute the average gender ratio separately for the two parts of the data set. Doing this for the sample data shows an average for the east of 94.65 males per 100 females and for the west of 98.46. This is not really a conclusive answer to the question, however.

6.4.2 Testing a Spatial Model

Perhaps a better way to state the final question in Section 6.4.1 is, Is there a statistical relationship between longitude and the gender ratio? If we had to express this mathematically, we could say that the gender ratio S is a function $f(\ )$ of longitude, or

$$S = f(\lambda)$$

The simplest form that this relationship might take is a linear relationship. You may remember from high school math the formula for a straight line, $y = mx + c$. The y is called the *dependent variable*, because its value depends on what is computed on the right-hand side of the equation and because it is the one we want to predict. The *independent variable*, x, is the one for which we have the data. The value m is the slope of the line, in this case the rate of increase or decrease if the value is negative, of the number of females per 100 males as we move through a degree of longitude. Finally, c is called the *intercept*. It is the value of the gender ratio at the place where longitude is zero. More about this later. In terms of our gender ratio data,

$$S = m\lambda + c \tag{6.1}$$

The first problem is how we select a single longitude to represent each state. We visited this problem earlier in this chapter in the context of Rugby, North Dakota. In this case we simply select a point placed by the highly subjective "eyeball" method—that is, the longitude of a point placed at the visually determined center of the state. Actually, the numbers selected were those used in a mapping program that drew circles to represent state values on a proportional circle map.

Now we can build a geographic database with only two attributes, the two we seek to relate, namely *gender ratio* and *longitude*. These two attributes are mapped in Figure 6.8. With only two variables, we can generate a scatter plot. The space of the scatter plot is "attribute" space, not geographic space. If there were a clear relationship as a straight line, the dots (each one a state) should line up along a sloping line. Take a look at Figure 6.9. Is there such a relationship? Just as we analyzed variation of a single attribute and then two attributes at once in Section 6.2, so also can we analyze variance for the two attributes here. We can use a method called *least squares* to model the scatter with a straight line.

Least squares calculates the variance between the two attributes (multiplied together) as a proportion of the variance in the independent variable. In each case, the total variance is computed as the sum of each attribute's values divided by the number of records (the mean), and this mean is subtracted from each of the attribute values before multiplying. Multiplying the two deviations from the means gives the cross-variance, and this value is divided by the regular squared variance for the independent variable.

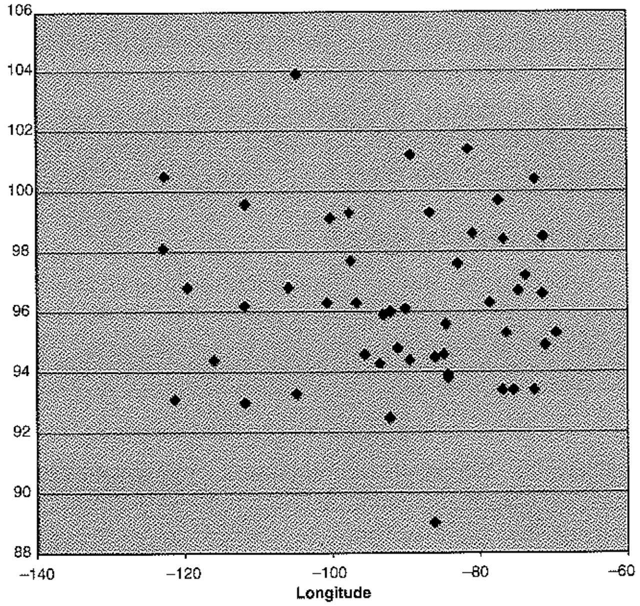


FIGURE 6.9: Scatter plot of Gender ratio data from Table 6.4. Gender ratio (Males per 100 females) is on the y axis, longitude of state capitals on the x axis.

We can use a formula to find values for m and c that minimize the sum of the squared variance. This gives the best-fit straight line through the data. From this we calculate that the formula linking the gender ratio and longitude has the form

$$S = -0.1438\lambda + 83.285 \tag{6.2}$$

The values predicted by this spatial model are mapped in the upper right-hand window in Figure 6.8. The strength of the statistical relationship is given by Pearson's correlation coefficient. The proportion of variance that the relationship "explains," a value called the *coefficient of determination* or r-squared, as a percentage is 61.8%. Obviously, the fit could be better. There must be a part of the geographic relationship that we missed.

6.4.3 Residual Mapping

A common way of seeking a deeper understanding of a spatial relationship is to examine the amount that each record deviates from the current model under analysis. For every record, if we plug the value for the independent variable (x) into the linear equation (6.2), we end up with an amount above or below the line in the up-down (y or dependent variable) direction on the scatter diagram. If we add all these together, they sum to zero, just as when we examined the deviations from the mean in the attribute description in Section 6.1.

These amounts are called *residual*s. Each record has a residual, just as each record has a geographic extent. Again, we can use the compute command or its equivalent in our GIS or database manager to calculate the residual for each state. It is the actual gender ratio for the state with the model value subtracted from it. Then we can use the GIS directly to make a map of the residuals. Just such a map is shown in Figure 6.8 in the lower right window.

This map takes a little interpretation but is a very powerful analytical tool. What the map of residuals shows is the unexplained variation in the simple linear model of gender ratio shown in the same figure on the lower left, and expressed in equation (6.2). Units are males per 100 females. Values in the range -1 to $+1$ mean that the model fits pretty well. High positive residuals are large underestimates, and high negative residuals are overestimates. As an examination of the map shows, by far the highest negative residual from the model is the District of Columbia. Thus the model seriously overestimated the number of males per hundred females in Washington, D.C. and there are far more females in Washington, D.C. than would be expected. The highest positive residuals are in the states of Nevada, New Hampshire, Vermont, Wyoming, and Colorado. In these areas, the model underestimated the gender ratio, i.e., there were more males than expected.

Regionally, the Deep South, the Pacific Northwest, and the "rust-belt" states seem to have negative residuals, meaning that the linear model overestimated their gender ratio, suggesting that they had more females than would be expected. It may be that these states have older populations and that women tend to live longer. States with more positive residuals (underestimates of the gender ratio, implying more males and fewer females in the population than the model predicts) are more scattered, with a dominance in the Mid-West and Mountain states. If we had a more complex linear model, involving both latitude and longitude, perhaps we could explain far more of the variation in the gender ratio. Conversely, some states seem to be exceptional: Nevada, for example. Should these states perhaps be investigated in more detail or at a finer scale? And what would happen if the linear model were applied for other time periods or other countries?

The example we have followed here has a few lessons to offer. First, spatial analysis follows the same path as much of scientific inquiry. We begin by displaying the attributes that we are interested in and looking at their aspatial (e.g., the histogram) and spatial (the map) characteristics. We try to see whether geographic properties have influenced the form of the distribution and if they are able to explain it. Then we formulate a model of the geographic relationship. In the preceding case this was a simple linear model relating the gender ratio and longitude.

We then formulate a way to test the model, usually a measure of goodness of fit between the model and the data that we have. In statistical terms, we suggest a hypothesis about the relationship, propose a counter or null hypothesis, and then devise a test to accept or reject the hypothesis based on the results. This often involves a probability from the normal distribution; in statistics, cutoff probabilities such as 95% or 99% are used to accept or reject a hypothesis.

Spatial analysis then goes further. We seek to explain geographically why the model does or does not fit. If the fit is inadequate, we can choose another model, change the geographic scope of the problem (e.g., the scale or the extent), or expand the model to include more attributes, that is, build a more complex model. Good science dictates that a simple model is preferred over a complex model, but that when a complex model explains a data set successfully, it is acceptable.

6.4.4 Prediction

The final stage of a model's use is to predict rather than to explain. Ideally, the geographic properties themselves have some explanatory basis and can point to a process in action. For example, if we analyzed a disease and found a high concentration in a single district, we could speculate that there was a clustered distribution with a single "source" and an outward diffusion from the source. Proof of this model would be to find a sample of people who had the disease and to show that they all contracted the disease in the single district. Prediction would then follow. An all-out attack on the disease in the one district would be the best strategy to eliminate the disease.

We can return to the gender ratio example to take a closer look at prediction. We can hypothesize that women live longer in the East and account for the linear relationship with longitude. This would obviously be extremely difficult to test by experiment. We could easily, however, use different data to examine the model. We could use county gender ratio data to see if the same model holds at different scales. We could use data for Mexico and the Canadian provinces to see whether the relationship carried over the edges of our geographic extent. We could also take the model to its limits, a very common tool in physics, math, and chemistry.

For example, our model uses longitude as the independent variable. What would the model predict as a gender ratio at the prime meridian, say for England and West Africa? What about for the middle of the Pacific Ocean at the 180th meridian? Clearly, the model has geographic constraints upon its predictive powers, and these can be tested and described. If the model is geographically invariant, this too is of great interest. Virtually every phenomenon on earth, even gravity, varies in some way over geographic space. If this were not the case, geography would not exist as a discipline, and GIS would be no more powerful than a spreadsheet program! There is value in mapping data even if no spatial analysis is intended, but merely as a visual description of a geographic pattern.

As GIS use becomes more commonplace, and as geographic explanations become of more value in managing resources, more of the geographic analyst's time will be spent in searching GIS data for spatial relationships. This geographic detective work can go a long way toward dealing with all the phenomena that have yet to be examined rigorously. This explains why GIS is being so rapidly accepted in disciplines such as archeology, demography, epidemiology, and marketing. In these fields, GIS allows the scientist to look for relationships that could not be seen without the lens of cartography and the integrative nature of GISs turned toward information and data like a high-powered telescope. The eye-opening moment of seeing the data visually for the first time is a common GIS convert's experience. Just as the explorers of the last century mapped out America and the world, so the GIS experts of today are mapping out new geographic worlds, invisible at the surface, but visible and crystal clear to the right tool controlled by the right vision.

In the search for spatial relationships, the GIS analyst is largely alone, however. The tools for searching are only now being integrated into GIS. In cartography, a process called the *design loop* is often used in map design. The map is generated to the specifications, and then the tools of digital cartography are used to make slow incremental improvements until the optimum design is reached. Obviously, not all of the improvements lead to success. Many serve to show what not to do; nevertheless, trial and error are important steps in the map's improvement.

Such a process is often used in GIS. A typical analysis is the result of extensive data collection, geocoding, data structuring, data retrieval, and map display, followed

ı. Ideally, the geographic
ıt to a process in action.
ation in a single district,
a single "source" and an
be to find a sample of
the disease in the single
isease in the one district

r look at prediction. We
or the linear relationship
test by experiment. We
l. We could use county
s. We could use data for
ip carried over the edges
its, a very common tool

variable. What would the
ngland and West Africa?
meridian? Clearly, the
l these can be tested and
great interest. Virtually
ver geographic space. If
e, and GIS would be no
napping data even if no
f a geographic pattern.
ic explanations become
lyst's time will be spent
detective work can go a
be examined rigorously.
nes such as archeology,
lows the scientist to look
aphy and the integrative
powered telescope. The
ime is a common GIS
apped out America and
graphic worlds, invisible
lled by the right vision.
largely alone, however.
n cartography, a process
generated to the spec-
make slow incremental
not all of the improve-
ertheless, trial and error

the result of extensive
map display, followed

by the sorts of description and analysis highlighted here. The next steps, prediction, explanation, and the use of these in decision making and planning, are the ultimate goal. Without seeing spatial relationships, however, these last prizes will be lost, and the GIS will not have reached its potential as an information management tool.

The GIS spatial relationship search loop consists of the steps outlined above, data assembly, preview, hypothesis design, hypothesis testing, modeling, geographic explanation, prediction, and examining the limitations of the model. It is the automated nature of the GIS and its flexibility that allows effective use of trial and error. In this process the GIS components most used are the database manager, for selecting and reorganizing attributes; the map display module, for display of intermediate and final analyses; and the computational or statistical tools available as part of the GIS. Some of these are sophisticated, although in many GIS packages they are rather simplistic.

One capability now being added to GIS is that of allowing examination of how distributions and relationships vary over time. Time is not particularly simple to integrate into a GIS, because each attribute data set and its maps are best interpreted as a single snap-shot in time. Nevertheless, both the attributes and the map are under constant change, and the geographic phenomena they represent are indeed very dynamic. Even an apparently stable attribute such as terrain is affected by mining, erosion, and volcanoes. Human systems are virtually all in a constant state of change.

We can simply compare two time periods for which we have data. Many human and social data are collected only once a decade, meaning that changes happening more rapidly than this will not be seen. Comparing two time periods allows only a single measurement, by value or map, of change. We can map, for example, on a satellite image all those areas that have changed between the time periods shown in the two images, assuming the same geographic extent at similar map scales. This can give us a direction of change, the addition or loss of wetlands, for example, and the amount of addition or loss, but not a rate of loss or gain. To measure a rate of change needs a minimum of three images or maps.

The most effective tool for time-sensitive geographic distributions is animation (Peterson, 1995). Animation allows the GIS interpreter to see changes as they take place. It is a key part of scientific visualization as far as a GIS is concerned, because usually more can be learned by examining the dynamics of a geographic system than its form. Imagine seeing glimpses of a chess game, for example, at only three stages in a game. The forms remain largely unchanged, although some pieces have moved around and disappeared from the board.

As we get more "frames" in the sequence, we arrive at the stage where every move is visible and the rules of movement of chess pieces become discernible. Finally, a great number of frames allow us to see not only the rules but the players themselves, their strategy, and the drama of the game. Just as there is an appropriate geographic resolution for map data, so is there a suitable time resolution. Like enlargement and reduction in space, time can be made slower or faster than "real" time to reach this appropriate resolution.

Figure 6.10 shows a few individual frames from an animation built with a GIS, in this case the historical growth of the Washington, DC, area. As is obvious, a static textbook cannot bring across the dynamic nature of the sequence. For this, use the MPEG sequence on the World Wide Web at `http://edcwww2.cr.usgs.gov/umap/umap.html`.

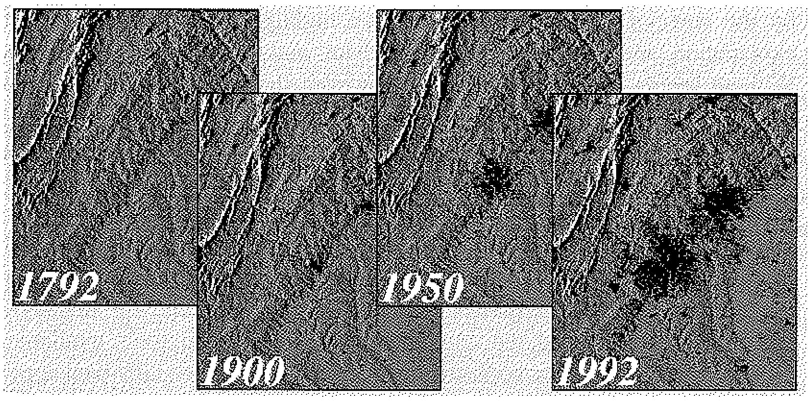


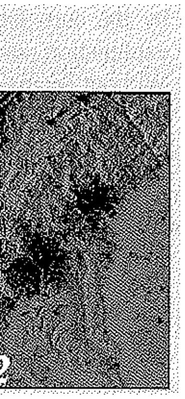
FIGURE 6.10: Frames from an animation of historical urban growth in the Washington, D.C. to Balitimore metropolitan area. Images by William Acevedo, USGS. Used with Permission.

6.4.5 Map Overlay Example

One of the oldest analytical methods used in GIS is map overlay. Map overlay is the set of procedures by which maps with different themes are brought into geometric and scale alignment so that their information can be cross referenced and used to createè more complex themes. We have met the method already several times, and should recall that the maps to be overlain must be of the same spatial extent, on the same map projection and datum, be at comparable granularity (that is, the spatial units, whether pixels or polygons, should be about the same average size), and if the layers are to be used with map algebra, at the same raster grid size and resolution.

The power of the GIS is in handling the geometry of the overlay process. Handling and preparing the themes is up to the GIS analyst. Under the simplest possible configuration, GIS layers are all converted to binary maps, and an overlay then sifts the map space to leave open the areas that satisfy the selection criteria in use. This is the case in the simple overlay analysis we met in the early chapters, and duplicates in the GIS methods that were worked out using transparent overlay maps and blacked out areas on the transparencies. Many of these methods date back to the turn of the twentieth century.

One means of map overlay is to intersect all of the layers involved to generate a set of most common geographic units. In map algebra, the raster plays this role. The attributes are then inherited or passed down to subsetted areas, and the attribute table gets longer and longer as more and more units are created. We have already seen the many problems with vector map overlay, including sliver polygons. Blind map overlay will happily assign attributes to very small sliver polygons, and use them in further analysis. A solution to this problem is to first process each layer to reduce the number of solution classes that will find their way into the final map. A selective query from each layer is one simple way to do this. A third overlay method is to find some common unit into which all values can be transformed. One GIS project that the author worked on solved the apparent incompatibility of the GIS layers for an ocean GIS by converting all of the themes into dollars, and adding them together to yield a composite. This is



ashington, D.C. to Balitimore

ay. Map overlay is the
ght into geometric and
and used to create more
, and should recall that
e same map projection
nits, whether pixels or
ers are to be used with

overlay process. Han-
r the simplest possible
d an overlay then sifts
criteria in use. This is
pters, and duplicates in
maps and blacked out
he turn of the twentieth

involved to generate a
er plays this role. The
and the attribute table
have already seen the
ons. Blind map overlay
d use them in further
r to reduce the number
A selective query from
s to find some common
that the author worked
cean GIS by converting
d a composite. This is

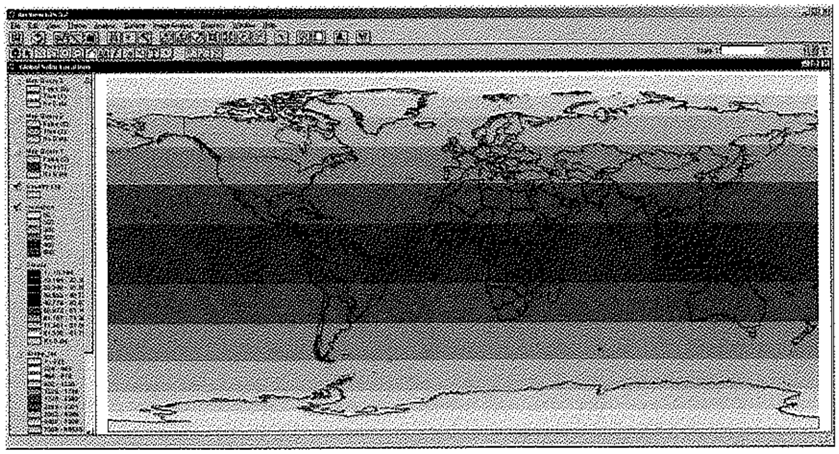


FIGURE 6.11: ArcView GIS layer of global insolation. Derived from a figure in *Geosystems*, by Robert Christopherson.

not always possible, of course, and far more common is to weight the overlays involved by some prechosen set of values reflecting the relative, not the absolute, importance of the layers.

For example, Figure 6.11 shows a world map of the factors considered for large scale siting of global solar power production. Information was searched for on the World Wide Web, downloaded and then processed into ESRI's ArcView so that map layers could be registered and converted to a common map projection using ArcView's projection extension. The themes used were those thought to relate to the possible supply and demand for solar-generated power. On the supply side, layers were the maps of incoming solar radiation, the expected average cloud cover, and global topography. On the demand side, a global population layer was used to buffer the solution set and limit it to those areas located close to major population agglomerations.

The overlay exercise consisted of creating a query to the GIS that essentially converted each of the layers into a binary map, a thresholding operation. The query requested insolation levels greater than 200 Watts per square meter, three levels of cloud coverage on average less than 65, 70, and 75%, and elevations lower than 5000 feet (1524 m), since although incoming solar power increases as the atmosphere becomes less dense with height, large scale construction favors lower and flat topography (Figures 6.12 and 6.13). These were combined by selecting those areas with population densities greater than 50 people per square kilometer (Figure 6.14). The resultant map shows that largest contiguous areas suitable for large scale solar power generation to include the Southwestern United States, Northern Chile, South Africa, the edges of the African Sahara, the Arabian peninsula, and Pakistan (Figure 6.15).

Having done the analysis with three different thresholds as far as cloud cover is concerned, two issues related to overlay analysis are shown. First, the different criteria are subjective, and need to be weighted to reflect their relative importance in the GIS solution. For example, it may be that total insolation is far more, say 10 times, more important than elevation in deciding where solar power generation is located. This can

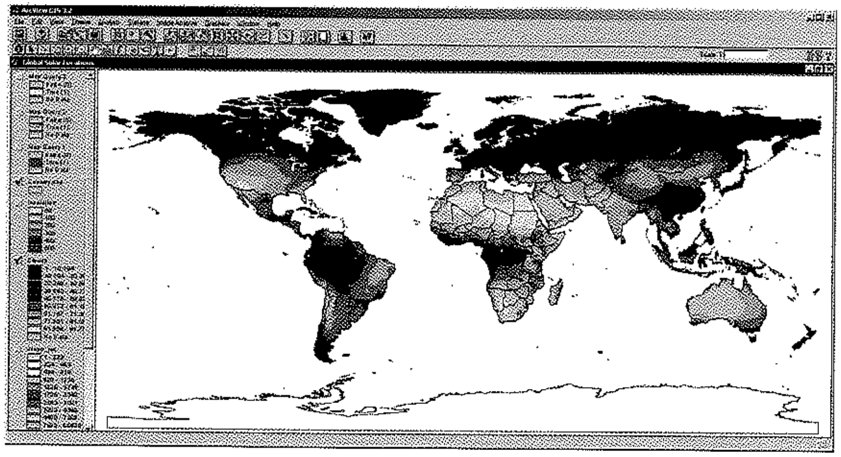


FIGURE 6.12: GIS layer of global average cloud cover. Source: UNEP Grid.

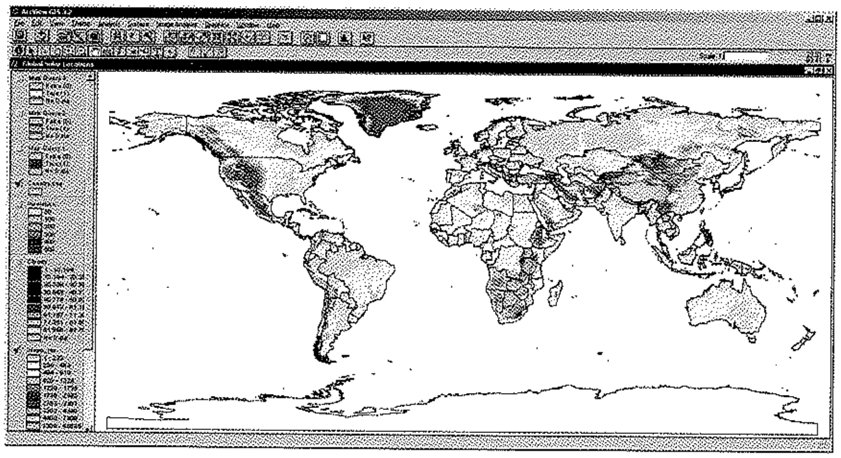


FIGURE 6.13: GIS layer of global topography. Source: Data set GTOPO30, United States Geological Survey, EROS Data Center.

be accommodated by multiplying the weights by the binary layers, each factored so that the sum of the weights is one, and then summing them across the layers. The final map will reflect the critical factors and their importance. However, selecting the weights can be a very complex process.

Second, in this case the solution area is highly influenced by only one of the layers, the cloud cover. This is why three levels of cloud cover were used in the final map. This "most sensitive" or critical layer in the map overlay process often marginalizes the contribution of other layers, perhaps even eliminating them from the final solution space altogether. This layer sensitivity is especially important to understand. Often a small amount of testing can reveal the critical layer. Some research suggests treating the layers as fuzzy, combining the factors together as a smooth field and creating margins of error on the solution map. Another important factor that affects analysis

rce: UNEP Grid.

Jnited States Geological Survey,

ers, each factored so that
the layers. The final map
selecting the weights can

d by only one of the lay-
ere used in the final map.
ocess often marginalizes
em from the final solu-
tant to understand. Often
research suggests treat-
mooth field and creating
ctor that affects analysis

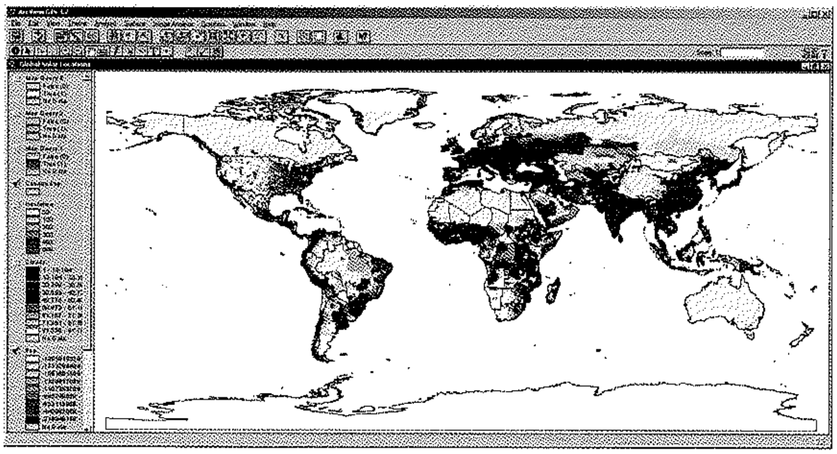
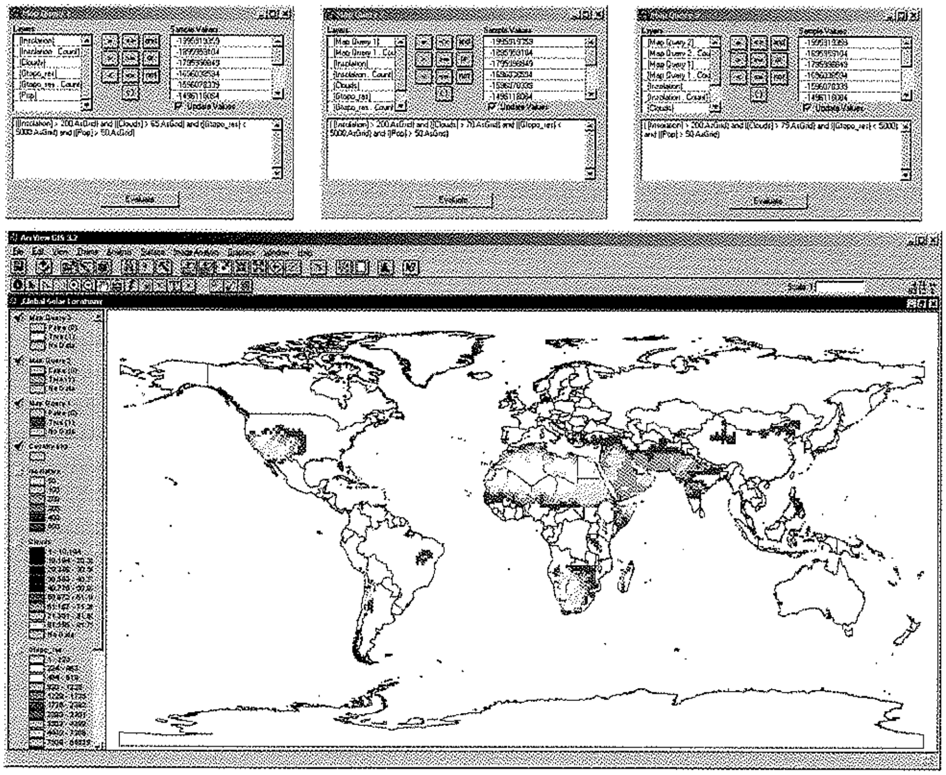


FIGURE 6.14: GIS layer of global population density. Source: NCGIA.



([Insolation] > 200.AsGrid) and ([Clouds] > 65.AsGrid) and ([Gtopo_res] < 5000.AsGrid) and ([Pop] > 50.AsGrid)

([Insolation] > 200.AsGrid) and ([Clouds] > 70.AsGrid) and ([Gtopo_res] < 5000.AsGrid) and ([Pop] > 50.AsGrid)

([Insolation] > 200.AsGrid) and ([Clouds] > 75.AsGrid) and ([Gtopo_res] < 5000) and ([Pop] > 50.AsGrid)

FIGURE 6.15: Solution map for the map overlay example. Global suitability for solar power production. Above the map are the three queries that generated the increasingly narrow solution sets. Figures 6.11 to 6.15 by Jeff Hemphill and Westerly Miller.

is how the error, for example caused by different generalization on each input layer, impacts the final decision by propagating through the overlay analysis and influencing the result.

Map overlay remains as one of the most common forms of GIS analysis. With the use of buffers and distance transforms, some very sophisticated analysis can be done. This method is extensively applied in planning, but is also increasingly used in all GIS applications, for fire modeling to habitat suitability mapping.

6.4.6 GIS and Spatial Analysis Tools

In the early days of GIS, much criticism was made of the fact that GIS software rarely came with any true analytical options. As we have seen, the basic tools of description are those of arithmetic and statistics, and the tools of modeling involve allowing the encoding of a model or formula into the system. Omitted here have also been models based entirely on the geographic distribution. Many models work on network flows, dispersion in two- or three dimensional space, hierarchical diffusion, or probabilistic models based on weights determined by buffers, and so on. This sort of model is manageable in a GIS using the tools of retrieval: overlay, buffering, and the application of spatial operators. Even a simple model, however, can become a quite lengthy sequence of steps for the GIS's user interface.

Almost all GIS packages allow operations to be bundled together as macros or as sequences of operations as part of a model, such as in the GISMO options. Although this goes a long way toward routine analysis, exploratory GIS data analysis is still something of an art. Many operations can be performed in the database manager only, and often GIS users move the data from the database manager into a standard statistical package such as SAS (Statistical Analysis System) or SPSS (Statistical Package for the Social Sciences) for analysis. One GIS (Arc/Info) offers a direct link to another statistical package (S-Plus) as an option.

Most GIS analysts use statistical and GIS tools in tandem during the analysis stage of GIS operation. The ability to produce nonspatial graphics—for example, a scatter plot or a histogram—is often far easier this way. Given the broad acceptance of statistical packages, and the large number of scientists and others trained in and familiar with their use, a compromise solution seems best. GIS packages can avoid duplicating the many functions necessary for statistical analysis by making two-way data movement between GIS and statistical software easy.

In summary, one of the greatest strengths of a GIS is that it can place real-world data into an organizational framework that allows numerical and statistical description and permits logical extension into modeling, analysis, and prediction. This important step, along with examining and thinking about one's data, is the bridge to understanding data geographically.

This understanding is enhanced by GIS because many phenomena simply cannot be understood, and certainly cannot be predicted, without an understanding of the geographic forces at work and their expression among the map's features as imprints of the principal geographic properties. Unfortunately, most GIS packages have contained only rudimentary tools for spatial analysis. However, GIS practitioners have filled the gaps with standard statistical software, and great strides are now being made as GIS contributes to the new models that are resulting in a host of different applications beyond the traditional scope of geography.

6.5 STUDY

6.5.1

CHAP

Descr

Statis

Spatia

on on each input layer,
analysis and influencing

f GIS analysis. With the
d analysis can be done.
easingly used in all GIS

that GIS software rarely
c tools of description are
ve allowing the encoding
en models based entirely
ws, dispersion in two- or
nodels based on weights
ageable in a GIS using
f spatial operators. Even
e of steps for the GIS's

together as macros or as
O options. Although this
nalysis is still something
ager only, and often GIS
atistical package such as
for the Social Sciences)
atistical package (S-Plus)

during the analysis stage
or example, a scatter plot
acceptance of statistical
n and familiar with their
id duplicating the many
data movement between

t it can place real-world
nd statistical description
ediction. This important
bridge to understanding

ienomena simply cannot
nderstanding of the geo-
eatures as imprints of the
ges have contained only
ners have filled the gaps
being made as GIS con-
rent applications beyond

6.5 STUDY GUIDE

6.5.1 Summary

CHAPTER 6: Why Is It There?

Describing Attributes (6.1)

- **GIS data description answers the question: Where? GIS data analysis answers the question: Why is it there?**
- **GIS data description is different from statistics because the results can be placed onto a map for visual analysis.**

Statistical Analysis (6.2)

- **The extremes of an attribute are the highest and lowest values, and the range is the difference between them in the units of the attribute.**
- **A histogram is a two-dimensional plot of attribute values grouped by magnitude and the frequency of records in that group, shown as a variable-length bar.**
- **For a large number of records distributed with random errors in their measurement, the histogram resembles a bell curve and is symmetrical about the mean.**
- **The mean is the sum of attribute values across all records, divided by the number of records. It is a representative value, and for measurements with normally distributed error, converges on the true reading.**
- **A value lacking sufficient data for computation is called a missing value.**
- **Accuracy is determined by testing measurements against an independent source of higher fidelity and reliability.**
- **The standard deviation is the average amount by which record values differ from the mean.**
- **The total variance is the sum of each record with its mean subtracted and then multiplied by itself.**
- **The standard deviation is the square root of the variance divided by the number of records less one.**
- **A sample is a set of measurements taken from a larger group or population. Sample means and variances can serve as estimates for their populations.**
- **The standard deviation in a spatial sense is a good descriptor of the accuracy of measurements.**
- **A mathematical version of the normal distribution can be used to compute probabilities associated with measurements with known means and standard deviations.**
- **A test of means can establish whether two samples from a population are different from each other, or whether the different measures they have are the result of random variation.**

Spatial Description (6.3)

- **For coordinates, data extremes define the two corners of a bounding rectangle.**
- **For coordinates, the means and standard deviations correspond to the mean center and the standard distance, both of which are good descriptors of spatial properties.**

- A centroid is any point chosen to represent a higher-dimension geographic feature, of which the mean center is only one choice.
- The standard distance for a set of point spatial measurements is the expected spatial error.
- Descriptions of geographic properties such as shape, pattern, and distribution are often verbal, but quantitative measures can be devised, although few are computed by GIS.
- GIS statistical computations are most often done using retrieval options such as buffer and spread, or by manipulating attributes with arithmetic commands.

Spatial Analysis (6.4)

- Geographic inquiry examines the relationships between geographic features collectively to help describe and understand the real-world phenomena that the map represents.
- Spatial analysis compares maps, investigates variation over space, and predicts future or unknown maps.
- Many GIS systems have to be coaxed to generate a full set of spatial statistics.
- A linear relationship is a predictable straight-line link between the values of a dependent and an independent variable. It is a simple model of the relationship.
- A linear relation can be tested for goodness-of-fit with least-squares methods. The coefficient of determination r-squared is a measure of the degree of fit, and the amount of variance explained.
- Differences between observed values of the dependent variable and those predicted by a model are called residuals.
- A GIS allows residuals to be mapped and examined for spatial patterns.
- A model helps explanation and prediction after the GIS analysis.
- A model should be simple, should explain what it represents, and should be examined at the limits before use.
- Overlay analysis is a very common form of GIS analysis.
- Overlay analysis requires maps on a common geometry with compatible themes.
- Themes can be weighted to reflect their relative importance.
- Often one theme dominates in controlling the final solution set to an overlay problem.

Searching for Spatial Relationships (6.5)

- Tools for searching out spatial relationships and for modeling are only lately being integrated into GIS.
- Statistical and spatial analytical tools are also only now being integrated into GIS, and many people use separate software systems outside the GIS.
- Real geographic phenomena are dynamic, but GISs have been mostly static. Time-slice and animation methods can help in visualizing and analyzing spatial trends.
- GIS places real-world data into an organizational framework that allows numerical description and lets the analyst model, analyze, and predict with both the map and the attribute data.

r-dimension geographic

rements is the expected

attern, and distribution
vised, although few are

g retrieval options such
h arithmetic commands.

een geographic features
-world phenomena that

over space, and predicts

l set of spatial statistics.
ink between the values
a simple model of the

n least-squares methods.
ure of the degree of fit,

variable and those pre-

or spatial patterns.
IS analysis.
presents, and should be

ysis.
with compatible themes.
rtance.
lution set to an overlay

nodeling are only lately

w being integrated into
outside the GIS.
have been mostly static.
zing and analyzing spa-

framework that allows
alyze, and predict with

6.5.2 Study Questions

Describing Attributes

Separate all of the readings in Table 6.1 taken east of Santa Barbara. Does this group represent a statistically different location and elevation from the remainder of the readings? Why or why not?

Write out step-by-step instructions for a child to calculate a mean and a median from a list of 10 numbers. Modify the instructions for 11 numbers. Write a one-paragraph explanation of what the numbers mean.

Spatial Analysis

Why is spatial analysis different, though related to, statistical analysis? Give examples.

Copy from a map such as a 1 : 24,000 series USGS map a group of objects, including a set of points with known elevations, some rivers, and forested areas. List as many measurements as you can devise to characterize the basic geographical properties of each feature, classified by dimension. Which measures are the easiest to calculate and why?

Draw a flow diagram of the stages of scientific inquiry surrounding analysis in a GIS. What is necessary before analysis can be conducted? What does a successful analysis lead to? What might prevent the ease of movement through the flow in the diagram?

Why have statistical and spatial analytical methods not been incorporated into all GIS packages?

6.6 EXERCISES

1. *For a GIS data set of your choosing, use only the tools available within the system to compute the extremes, the mean, and the standard deviation for all the attributes in the data set. What problems do you encounter along the way?*
2. *How might the length of a line and the area of a polygon be calculated (a) in a vector GIS and (b) in a raster GIS? Why might they be expected to give different results?*
3. *Design a model that might account for the risk of wildfire for a GIS data set consisting of layers for vegetation type and condition, soils, streams, topography, and wind direction. How might the model be tested?*
4. *Use a GIS to overlay a map of topography with polygonal districts such as counties. Using any method available, compute and then map the variance or the standard deviation of the elevation values within each district. Explain the distribution on the map.*
5. *Repeat the gender ratio example in the chapter for data on mean annual precipitation by state. Does the spatial analysis result in a stronger relationship? Why might this be?*
6. *Trace a feature from a map, or select a single polygon feature within a GIS. Choose as complex a feature as you can find. Using the GIS's capabilities, locate as many centroids by as many methods as you can devise. Do any of these points fall outside the polygon? Which methods give similar results?*

6.7 REFERENCES

Campbell, J. (1993) *Map Use and Analysis*, 2nd ed. Dubuque, IA: Wm. C. Brown.

Earickson, R. and Harlin, J. (1994) *Geographic Measurement and Quantitative Analysis*. New York: Macmillan.

Peterson, M. P. (1995) *Interactive and Animated Cartography*. Upper Saddle River, NJ: Prentice Hall.

The World Almanac and Book of Facts. New York: Pharos Books. Published annually.

Unwin, D. (1981) *Introductory Spatial Analysis*. London: Methuen.

6.8 KEY TERMS AND DEFINITIONS

analysis: The stage of scientific inquiry when data are examined and tested for structure in support of hypotheses.

attribute: An item for which data are collected and organized. A column in a table or data file.

bearing: An angular direction given in degrees from zero as north, clockwise to 360.

bell curve: A common term for the normal distribution.

bounding rectangle: The rectangle defined by a single feature or a collection of geographical features in coordinate space, and determined by the minimum and maximum coordinates in each of the two directions.

centroid: A point location at the center of a feature used to represent that feature.

compute command: In a database manager, a command allowing basic arithmetic on attributes or combinations of attributes, such as summation, multiplication, and subtraction.

converge: The eventual agreement of measurements on a single value.

data extremes: The highest and lowest values of an attribute, found by selecting the first and last records after sorting.

dependent variable: The variable on the left of the equals sign in a formula model, whose values are determined by the values of the other variables and constants.

difference of means: A statistical test to determine whether or not two samples differ from each other statistically.

error band: The width of a margin plus and minus one standard error of estimation, as measured about the mean.

expected error: One standard deviation in the units of measure.

goodness of fit: The statistical resemblance of real data to a model, expressed as strength or degree of fit of the model.

gradient: The constant of multiplication in a linear relationship, that is, the rate of increase of a straight line up or down. See also **slope**.

histogram: A graphic depiction of a sample of values for an attribute, shown as bars raised to the height of the frequency of records for each class or group of value within the attribute.

hypothesis: A supposition about data expressed in a manner to make it subject to statistical test.

independent variable: A variable on the right-hand side of the equation in a model, whose value can range independently of the other constants and variables.

intercept: The value of the dependent variable when the independent variable is zero.

least squares: A statistical method of fitting a model, based on minimizing the sum of the squared deviations between the data and the model estimates.

linear relationship: A straight-line relationship between two variables such that the value of the dependent variable is a gradient times the independent variable plus a constant.

mean: A representative value for an attribute, computed as the sum of the attribute values for all records divided by the number of records.