CHAPTER 1

Space–Time: The Next Frontier

This book is about the statistical analysis of data . . . *spatio-temporal* data. By this we mean data to which labels have been added showing where and when they were collected. Good science protocol calls for data records to include place and time of collection. Causation is the "holy grail" of Science, and hence to infer cause–effect relationships (i.e., "why") it is essential to keep track of "when"; a cause always precedes an effect. Keeping track of "where" recognizes the importance of knowing the "lay of the land"; and, quite simply, there would be no History without Geography.

We believe that in order to answer the "why" question, Science should address the "where" and "when" questions. To do that, spatio-temporal datasets are needed. However, spatial datasets that do not have a temporal dimension can occur in many areas of Science, from Archeology to Zoology. The spatial data may be from a "snapshot" in time (e.g., liver-cancer rates in U.S. counties in 2009), or they may be taken from a process that is not evolving in time (e.g., an iron-ore body in the Pilbara region of Australia). Sometimes, the temporal component has simply been discarded, and the same may have happened to the spatial component as well. Also, temporal datasets that do not have a spatial dimension are not unusual, for analogous reasons. For example, two time series, one of monthly mean carbon dioxide measurements from the Mauna Loa Observatory, Hawaii, and the other of monthly surface temperatures averaged across the globe, do not have a spatial dimension (for different reasons).

Spatio-Temporal Data

Spatio-temporal data were essential to the nomadic tribes of early civilization, who used them to return to seasonal hunting grounds. On a grander scale, datasets on location, weather, geology, plants, animals, and indigenous people were collected by early explorers seeking to map new lands and enrich their kings and queens. The conquistadors of Mesoamerica certainly did this for Spain.

Statistics for Spatio-Temporal Data, by Noel Cressie and Christopher K. Wikle

The indigeneous people also made their own maps of the Spanish conquest, in the form of a *lienzo*. A lienzo represents a type of historical cartography, a painting on panels of cloth that uses stylized symbols to tell the *history* of a *geographical* region. The *Lienzo de Quauhquechollan* is made up of 15 joined pieces of cotton cloth and is a map that tells the story, from 1527 to 1530, of the Spanish conquest of the region now known as Guatemala. It has been restored digitally in a major project by Exploraciones sobre la Historia at the Universidad Francisco Marroquín (UFM) in Guatemala City (see Figure 1.1). This story of the Spanish conquest in Guatemala is an illustration of complex spatio-temporal interactions. Reading the lienzo and understanding its correspondence with the geography of the region required deciphering; see Asselbergs (2008) for a complete description. The original lienzo dates from about 1530 and represents a spatio-temporal dataset that is almost 500 years old!

In a sense, we are all analyzers of spatial and temporal data. As we plan our futures (economically, socially, academically, etc.), we must take into account the present and seek guidance from the past. As we look at a map to plan a trip, we are letting its spatial abstraction guide us to our destination. The philosopher Ludwig Wittgenstein compared language to a city that has evolved over time (Wittgenstein, 1958): "Our language can be seen as an ancient city: A maze of little streets and squares, of old and new houses, and of houses with additions from various periods; and this surrounded by a multitude of new burroughs with straight and regular streets and uniform houses!"

Graphs of data indexed by time (time series) and remote-sensing images made up of radiances indexed by pixel location (spatial data) show variability at a glance. For example, Figure 1.2 shows the Missouri River *gage-height* levels during the 10-year period, 1988–1997, at Hermann, MO. Figure 1.3 shows two remotely sensed images of the river taken in September 1992, before a major flood event, and in September 1993, after the highest crest ever recorded at Hermann (36.97 ft on July 31, 1993). The top panel of Figure 1.3 shows the town of Gasconade in the middle of the scene, situated in the "V" where the Gasconade River joins the Missouri River; Gasconade is at mile 104.4 and eight miles downstream is the river town of Hermann, visible at the very bottom of the scenes. Notice the intensive agriculture in the river's flood plain in September 1992. The bottom panel of Figure 1.3 shows the same region, one year later, after the severe flooding in the summer of 1993. The inundation of Gasconade, the floodplain, and the environs of Hermann is stunning. There is a multiscale process behind all of this that involves where, when, and how much precipitation occurred upstream, the morphology of the watershed, microphysical soil properties that determine run-off, the U.S. Army Corps of Engineers' construction of levees upstream, and so on. However, by looking only in the spatial dimension, or only in the temporal dimension, we miss the dynamical evolution of the flood event as it progressed downstream. Spatio-temporal data on this portion of the Missouri River, which shows how the river got from "before" to "after," would be best illustrated with a movie, showing a temporal sequence of spatial images before, during, and after the flood.

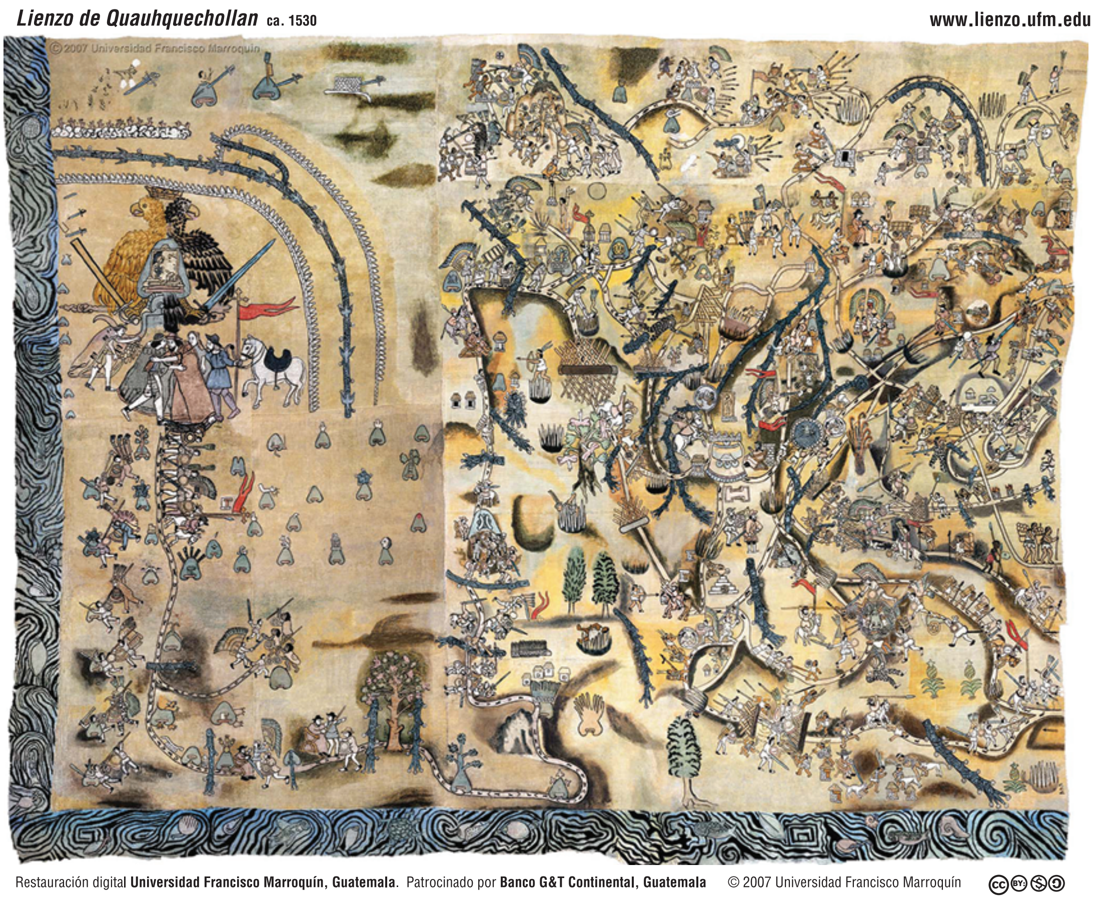

Figure 1.1 Digitally restored Lienzo de Quauhquechollan, whose actual dimensions are 2.45 m in height by 3.20 m in width. [Image is available under the Creative Commons license Attribution-Noncommercial-Share Alike © 2007 Universidad Francisco Marroquín.]

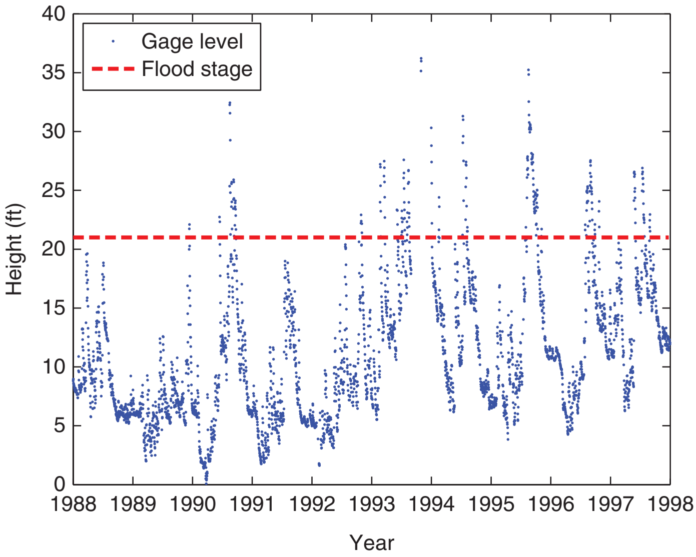

Figure 1.2 Time-series levels of gage height at Hermann, MO (mile 96.5 on the Missouri River) from January 1, 1988 through December 31, 1997. Flood stage is given by the horizontal dashed line. The highest recorded gage height in the 10-year period was 36.97 ft on July 31, 1993.

There is an important statistical characteristic of spatio-temporal data that is very common, namely that nearby (in space and time) observations tend to be more alike than those far apart. However, in the case of "competition," the opposite may happen (e.g., under big trees only small trees can grow), but the general conclusion is nevertheless that spatio-temporal data should *not* be modeled as being statistically independent. [Tobler, (1970) called this notion "the first law of Geography."] Even if spatio-temporal trends are used to capture the dependence at large scales, there is typically a cascade of smaller spatio-temporal scales for which a statistical model is needed to capture the dependence. Consequently, an assumption that spatio-temporal data follow the "independent and identically distributed" (*iid*) statistical paradigm should typically be avoided. Paradigms that incorporate dependence are needed: The time series models in Chapter 3 and the spatial process models in Chapter 4 give those paradigms for *temporal data* and *spatial data*, respectively. From Chapter 5 onwards, we are concerned directly with Statistics for *spatio-temporal data*.

Uncertainty and the Role of Statistics

Uncertainty is everywhere; as Benjamin Franklin famously said (Sparks, 1840), "In this world nothing can be said to be certain, except death and taxes." Not only is our world uncertain, our attempts to explain the world (i.e., Science) are uncertain. And our measurements of our (uncertain) world are uncertain. Statistics is the "Science of Uncertainty," and it offers a coherent approach to

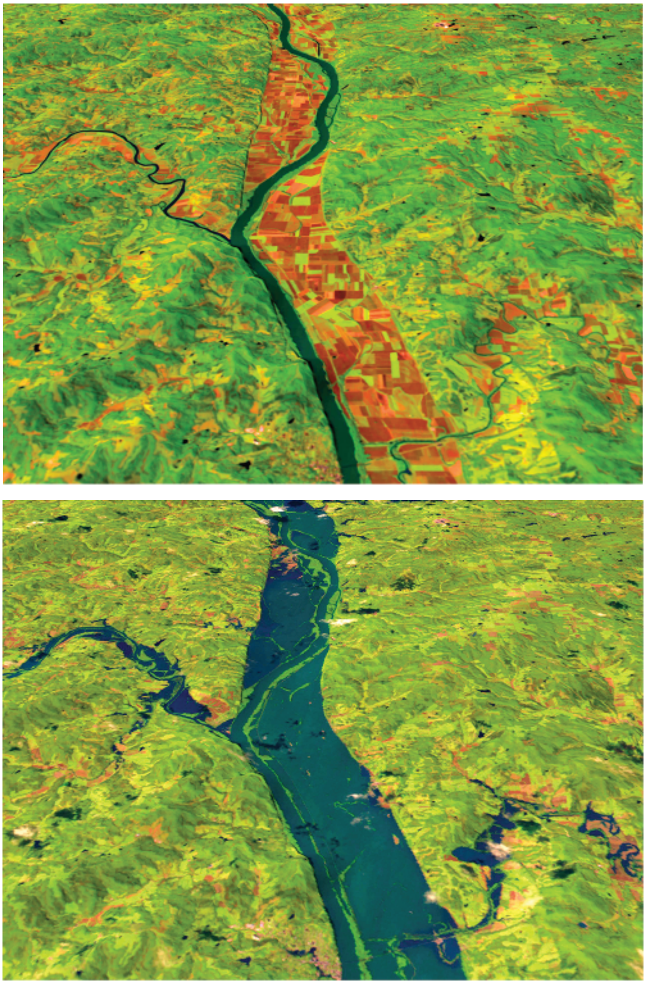

Figure 1.3 Images from NASA's Landsat Thematic Mapper. Each image shows a segment of the Missouri River near Hermann, MO (mile 96.5, at the bottom of the scene), and Gasconade, MO (mile 104.4, in the "V" in the middle of the scene). The river flows from west (top of the scene) to east (bottom of the scene). **Top panel:** September 1992, before a major flood event. **Bottom panel:** September 1993, after a record-breaking flood event in July 1993.

handling the sources of uncertainty referred to above. Indeed, in our work we use the term *Statistical Science* interchangeably with *Statistics* (with a "capital" S); we use *statistics* (with a "small" s) to refer to summaries of the data.

In most of this book, we shall express uncertainty through variability, but we note that other measures (e.g., entropy) could also be used. Just as the physical and biological sciences have the notions of mass balance and energy balance, Statistical Science has a notion of variability balance. The total variability is modeled with variability due to *measurement*, variability due to using a (more-or-less uncertain) *model* of how the world works, and variability due to uncertainty on *parameters* that control the measurement and model variabilities.

Although real-world systems may in principle be partially deterministic, our information is incomplete at each of the stages of observation, summarization, and inference, and thus our understanding is clouded by uncertainty. Consequently, by the time the inference stage is reached, the lack of certainty will influence how much knowledge we can gain from the data. Furthermore, if the dynamics of the system are nonlinear, the processes can exhibit *chaos* (Section 3.2.4), even though the theory is based on *deterministic* dynamical systems. (In Chapters 3 and 7, we show how model uncertainty in these systems naturally leads to *stochastic* dynamical systems that incorporate *system*, or *intrinsic,* noise.)

Data can hold so much potential, but they are an entropic collection of digits or bits unless they can be organized into a database. With the ability in a database to structure, search, filter, query, visualize, and summarize, the data begin to contain *information*. Some of this information comes from judicious use of statistics (i.e., summaries) with a "small s." Then, in going from information to *knowledge*, Science (and, with it, Statistics with a "capital S") takes over. This book makes contributions at all levels of the data–information–knowledge pyramid, but we generally stop short of the summit where knowledge is used to determine policy. The methodology we develop is poised to do so, and we believe that at the interface between Science, Statistics, and Policy there is an enormous need for (spatio-temporal) decision-making in the presence of uncertainty.

In this book, we approach the problem of "scientific understanding in the presence of uncertainty" from a probabilistic viewpoint, which allows us to build useful spatio-temporal statistical models and make scientific inferences for various spatial and temporal scales. Accounting for the uncertainty enables us to look for possible associations within and between variables in the system, with the potential for finding mechanisms that extend, modify, or even disprove a scientific theory.

Uncertainty and Data

Central to the observation, summarization, and inference (including prediction) of spatio-temporal processes are *data*. All data come bundled with error. In particular, along with the obvious errors associated with measuring, manipulating, and archiving, there are other errors, such as discrete spatial and temporal sampling of an inherently continuous system. Consequently, there are always scales of variability that are unresolvable and that will further "contaminate" the observations. For example, in Atmospheric Science, this is considered a form of "turbulence," and it corresponds to the well known aliasing problem in time series analysis (e.g., see Section 3.5.1; Chatfield, 1989, p. 126) and the microscale component of the "nugget effect" in geostatistics [e.g., see the introductory remarks to Chapter 4 and Cressie (1993, p. 59)].

Furthermore, spatio-temporal data are rarely sampled at spatial or temporal locations that are optimal for the analysis of a specific scientific problem. For instance, in environmental studies there is often a bias in data coverage toward areas where population density is large, and within a given area the coverage may be limited by cost. Thus, the location of a measuring site and its temporal sampling frequency may have very little to do with the underlying scientific mechanisms. A scientific study should include the *design* of data locations and sampling frequencies when framing questions, when choosing statistical-analysis techniques, and when interpreting results. This task is complicated, since the data are nearly always statistically dependent in space and time, and hence most of the traditional statistical methods taught in introductory statistics courses (which assume *iid* errors) do not apply or have to be modified.

Uncertainty and Models

Science attempts to *explain* the world in which we live, but that world is very complex. A model is a simplification of some well chosen aspects of the world, where the level of complexity often depends on the question being asked. Pragmatically, the goal of a model is to predict, and at the same time scientists want to incorporate their understanding of how the world works into their models. For example, the motion of a pendulum can be modeled using Newton's second law and the simple gravity pendulum that ignores the effect of friction and air resistance. The model predicts future locations of the pendulum quite well, with smaller-order modifications needed when the pendulum is used for precise time-keeping. Models that are scientifically meaningful, that predict well, and that are conceptually simple are generally preferred. An injudicious application of Occam's razor (or "the law of parsimony") might elevate simplicity over the other two criteria. For example, a statistical model based on correlational associations might be simpler than a model based on scientific theory. The way to bridge this divide is to focus on what is more or less certain in the scientific theory and use *scientific-statistical* relationships to characterize it.

Albert Einstein said: "It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience," at the Herbert Spencer Lecture delivered at Oxford University on June 10, 1933; see Einstein (1934). Much later, in the October 1977 issue of the *Reader's Digest*, it appears as if Einstein's quote was paraphrased to: "Everything should be made as simple as possible, but not simpler." Statistics and its models, including those involving scientific–statistical relationships, should not be spared from following this advice. Royle and Dorazio (2008, pp. 414–415) give a succinct discussion of this desire for conceptual simplicity in a model. As the data become more expansive, it is natural that they might suggest a more complex model. Clearly, there is a balance to be struck between too much simplicity, and hence failing to recognize an important signal in the data, and too much complexity, which results in a nonexistent signal being "discovered." One might call this desire for balance the *Goldilocks Principle* of modeling. (*Goldilocks and the Three Bears* is a nursery tale about a little girl's discovery of what is "just right.")

It is our belief that statistical models used for describing temporal variability in space should represent the variability dynamically. Models used in Physics, Chemistry, Biology, Economics, and so on, do this all the time with difference equations and differential equations to express the evolutionary mechanisms. Why should this change when the models beome statistical? Perhaps it is because there is often an alternative—for example, a model based on autocorrelations that describe the temporal dependence. However, this descriptive approach does not directly involve evolutionary mechanisms and, as a consequence, it pushes understanding of the Physics/Chemistry/Biology/Economics/etc. into the background. As has been discussed above, there is a way to have both, in the form of a scientific–statistical model that recognizes the dynamical scientific aspects of the phenomenon, with their *uncertainties* expressed through *statistical* models. Descriptive (correlational) statistical models do have a role to play when little is known about the etiology of the phenomenon; this approach is presented in Sections 6.1 and 6.2. Thereafter, this book adopts a dynamical approach to Statistics for spatio-temporal data.

Nearby Things Tend to Be More Alike...

A simple and sometimes effective forecast of tomorrow's weather is to use today's observed weather. This "persistence" forecast is based on observing large autocorrelations between successive days. Such dependence behavior in "nearby" temporal data is also seen in "nearby" spatial data, such as in studies of the environment. Statistics for spatio-temporal data presents the next frontier; this book steps forward into new territories and revisits old ones. It reviews and extends different aspects of statistical methodology based on spatio-temporal dependencies: exploratory data analysis, marginal/conditional models in discrete/continuous time, optimal inference (including parameter estimation and process prediction), model diagnostics and evaluation, and so forth. One fundamental scientific problem that arises is understanding the evolution of processes over time, particularly in environmental studies (e.g., the evolution of sea-ice coverage in the Arctic; changes in sea level; time trends in precipitation). Proper inference to determine if evolutionary components (natural or anthropogenic) are real requires a spatio-temporal *statistical* methodology.

The scientific method involves observation, inspiration, hypothesis generation, experimentation (to support or refute the current scientific hypothesis), inference, more inspiration, more hypothesis generation, and so forth. In a sense, everything begins with observation, but it is quickly apparent to a scientist that unless data are obtained in a more-or-less controlled manner (i.e., using an experimental design), proper inference can be difficult. This is the fundamental difference between "observation" and "experimentation." Understanding the role of dependencies when the data are spatial or temporal, or both, provides an important perspective on working with experimental data versus observational data.

Experimental Data

Earth's population is many billions, and the demand for sustenance is great and continuous. The planet's ability to produce food on a massive scale largely came from fundamental experiments in crop science in the early twentieth century. Fisher (1935) developed a statistical theory of experimental design, based on the three principles of blocking, randomization, and replication, for choosing high-yielding, insect-resistant crops adapted to local conditions. He developed a vocabulary that is used today in scientific experiments of all types: response (e.g., wheat yields), treatments (e.g., varieties of wheat), factors (e.g., soil type, field aspect, growing season), levels of factors (e.g., for the soil-type factor, the levels might be sand, gravel, silt, clay, peat), plot (experimental unit that receives a single treatment), block (collection of plots with the same factor/level combination), randomization (random assignment of treatments to plots), replication (number of responses per treatment), and so on.

Data from designed experiments, when analyzed appropriately, allow stronger (almost) causative inferences, which incubate further scientific inspiration and hypothesis generation, and so forth, through the cycle. In the right hands, and with a component of luck, this cycle leads to great breakthroughs [e.g., the discovery of penicillin in 1928 by Alexander Fleming; see, e.g., Hare (1970)]. Even small breakthroughs are bricks that are laid on the knowledge pyramid.

Space and time are fundamental factors of any experiment. For example, "soil type" is highly spatial and "growing season" is highly temporal. Protocol for any well designed experiment should involve recording the location and time at which each datum was collected, because so many factors (known or unknown) correlate with them. After the experiment has been performed, spatial and temporal information can be used as proxies for unknown, unaccounted-for factors that may later become "known" as the experiment proceeds. From this point of view, the natural place to put spatial and temporal effects in the statistical model is in the mean. But, there is an alternative . . . .

In R. A. Fisher's pathbreaking work on design of experiments in agricultural science, he wrote (Fisher, 1935, p. 66): "After choosing the area we usually have no guidance beyond the widely verified fact that patches in close proximity are commonly more alike, as judged by the yield of crops, than those which are further apart." Spatial variability, which to Fisher came in the form of plot-to-plot variability, is largely due to physical properties of the soil and environmental properties of the field. Fisher avoided the confounding of treatment effect with plot effect with the inspirational introduction of *randomization* into the scientific method. It was a brilliant insertion of *more* uncertainty into a place in the experiment where uncertainty abounds, leaving the more certain parts of the experiment intact. Fisher's idea has had an enormous effect on all our lives. For example, any medicine we have taken to treat our ailments and illnesses has gone through rigorous testing, to which the *randomized* clinical trial is central (where a "plot" is often the patient).

Randomization comes with a price. It allows valid inference on the treatments through a simple expression for the mean response, but the variances and covariances of the responses are affected too. Under randomization of the assignment of treatments to plots, the notions of "close proximity" and "far apart" have been hustled out the back door. Can we get spatial dependence back into the statistical analysis of responses, resulting in more efficient inferences for treatment effects? The answer to this is a resounding "Yes"; see the introduction to Chapter 4.

Observational Data

Organisms are born, live, reproduce, and die, but they can produce harmful by-products that may threaten their own well-being as well as the well-being of other organisms around them. (The species *Homo sapiens* is unique in many of its abilities, including its ability to have a major impact on all other organisms on Earth.) Variability within organisms can be large (e.g., within *H. sapiens*), as can variability between their environments. Thus, it can be very difficult to conduct controlled experiments on Earth's ecology and environment.

Observational data come from a "wilder side" of Science. The environment (such as climate, air and water quality, radioactive contamination, etc.) is a part of our lives that often will not submit to blocking, randomization, and replication. We cannot control when it rains, nor can we observe two Los Angeles, one with smog and one without. We *can* look for two like communities, one with contaminated water and one without; and we *can* look at health records before and after a toxic emission. However, any inference is tentative because the two factors, space and time, are not controlled for. Collecting samples from ambient air presents a philosophical problem because the parcel of air is unique when it passes the monitoring site; it evolves as the changes in air pressure move it around, and it will never come back to allow us the luxury of obtaining an independent, identically distributed observation. (If these observations are used to study the effect of air quality on human health, there is the further problem that the ambient air is not actually what individuals breathe in their homes or their workplaces; this introduces even more uncertainty into the study.)

In the environmental and life sciences, classical experimental design can struggle to keep up with the questions being asked, but they still need to be answered. And, as we have discussed just above, *uncertainty* is likely to be higher without experimental control. Thus, Statistical Science has a crucial role to play, although it does not fit neatly into the blocking–randomization–replication framework. Even when one is able to "block" the human subjects on age and sex, say, it may be that an unknown genetic factor will determine how a patient responds to a given treatment. (Personalized medicine has as one of its goals to make the unknown genetic factor known.) In epidemiological studies, controls may be randomly matched with cases, but the cases are in no way assigned randomly to neighborhoods. And, although duplicate chemical assays allow for assessment of measurement error in a study on stream pollution, replication of a water parcel from the stream is impossible. In such circumstances, Statistics is even more relevant, and we advocate that the scientific method invoke the principle of *expressing uncertainty through probabilities*.

In the environmental sciences, proximity in space and time is a particularly relevant factor. The word "environ" means "around" in French. While ecology is the study of organisms, the *environment* is the surroundings of organisms. "Nearby" is a relative notion, relative to the spatial and temporal scales of the phenomenon under study. For example, in the spatial case, a toxic-waste-disposal site may directly affect a neighborhood of a few square miles; a coal-burning power plant may directly affect a heavily populated region of many tens of square miles, and an increase in greenhouse gases will affect the whole planet. Clearly, a global effect is felt locally in many ways, from a longer growing season in Alberta, Canada, to a redistribution of beachfront property in Florida, USA. The point we wish to make here is that a quantity like global mean temperature is a largely uninformative summary of how daily lives of a community will be affected by a warmer planet, which means that environmental studies of the globe must recognize the importance of *local* variability. Furthermore, how the spatial variability behaves dynamically (i.e., the spatio-temporal variability) is key to understanding the causes of global warming and what to do about it. Finally, we state the obvious, that political boundaries cannot hold back a one-meter rise in sea level; our environment is ultimately a global resource and its stewardship is an international responsibility.

Einsteinian Physics

Einstein's theory of relativity (e.g., Bergmann, 1976) demonstrated that space and time are interdependent and inseparable. In contrast, our book is almost exclusively concerned with phenomena that reside in a classical Newtonian framework (e.g., Giancoli, 1998). We include a brief discussion of space and time within Einstein's framework, to indicate that modifications would be needed for, say, spatio-temporal astronomical data.

Einstein proposed a "thought experiment," a version of which we now give. Think of a boxcar being pulled by a train traveling at velocity v, and place a source of light at the center of the moving boxcar. An observer on the train sees twin pulses of light arrive at the front and rear end of the boxcar, *simultaneously*. A stationary observer standing by the train tracks sees one pulse arrive at the rear end of the boxcar *before* its twin arrives at the front end. That is, the reference frame of the observer is extremely important to the temporal notions of simultaneity/before/after. What ties together space and time is movement (velocity) of the boxcar.

Einsteinian physics assumes that the velocity of light c, is a universal constant (which is approximately 3×10^5 km/s), regardless of the frame of reference. Thus, for *any* frame of reference, the distance traveled by a pulse of light is equal to the time taken to travel that distance multiplied by c. That this relationship holds under any spatio-temporal coordinate system means that for Einsteinian physics, space and time are inextricably linked. Other physical properties are modified too. The length of an object measured in the moving frame, moving with velocity v, is always smaller than or equal to the length of the object measured in the stationary frame, by a factor of $\{1 - (v/c)^2\}^{1/2}$. A similar factor shortens a time interval in a moving frame, leading to the famous conclusion that the crew of a spaceship flying near the speed of light would return in a few (of *their*) years to find that their generation on Earth had become old.

Einstein's theory of relativity is most certainly important for some phenomena, but in this book we shall stay within scales of space and time where the physical laws of Newton can be assumed. We work with a coordinate system that is a Cartesian product of three-dimensional space and one-dimensional time, while respecting the directionality of the temporal coordinate. Our models of spatio-temporal processes attempt to capture the complex statistical dependencies that can arise from the *evolution* of phenomena at many *spatial and temporal scales*.

Change-of-Support

The global/regional/local scales of spatio-temporal variability lead to a phenomenon we shall call *change-of-support.* In the spatial case, it is known as downscaling/upscaling, or the ecological effect, or the modifiable areal unit problem. It is in fact a manifestation of Simpson's paradox (Simpson, 1951). Simpson's paradox, which has a perfectly rational probabilistic explanation, essentially says the following: In a two-way cross-tabulation, the variables (A and B, say) can exhibit a positive statistical dependence, yet when a third variable (C, say) enters and expands the data to a three-way cross-tabulation, the statistical dependence between A and B can be negative for *each* value of C!

For example, consider the data reported in Charig et al. (1986) and discussed by Meng (2009), on the treatment of kidney stones. *Open surgery* had a success rate of 78%, not as good as the *ultrasound* treatment's success rate of 83%. However, for small stones (<2-cm mean diameter), the success rate for open surgery was 93% and that for ultrasound was 87%. That is, open surgery did better than ultrasound for small stones. Surely, for large stones (≥2-cm mean diameter), open surgery would do worse, to account for its inferior success rate based on the results given above for all stones (78% versus ultrasound's 83%). Not so! For large stones, the success rate for open surgery was 73% and that for ultrasound was 69%, *again in favor of open surgery*. This is a sober reminder to all scientists to respect the "lurking variable," manifested here as the size of a patient's kidney stone.

Similarly, in a temporal setting, a causal statistical model built at a 3-monthly scale may have little or no relevance to the mechanisms in play at the daily scale. Day trading on stock markets, based on economic relationships estimated from quarterly trade figures, would probably lead to financial ruin. In a spatial setting, regional climate data may warn, correctly, of a future drought in the Northwest United States (states of Washington and Oregon). However, local orographic effects may favor certain parts of the Willamette Valley in Oregon to the point where above-average rainfall is consistently received there. That is, rather than size of kidney stone, think of Simpson's paradox in terms of size of region (space) or length of period (time). See Cressie (1998a) for a discussion of change-of-support in a spatio-temporal setting.

As we have mentioned, aggregations over time are subject to the change-of-support effect, but there is less discussion of it in the time series literature because time series are often already downscaled to answer the questions of interest. In contrast, spatial aggregation is ubiquitous: In the United States, federal decisions (e.g., carbon "cap and trade") are made at a continental scale, state decisions (e.g., California's clean-air regulations) are made at a regional scale, and city-wide decisions (e.g., the water-conservation policy in the city of Tucson, Arizona) are made at a local scale. These decisions are based on data that come from a variety of spatial scales; however, an inappropriate statistical analysis that does not respect the change-of-support effect could lead to the adoption of inappropriate policies. Our goal in this book is to build spatio-temporal statistical models to *explain* the variability in observable phenomena. While change-of-support should always be respected, there is less of a chance it will cause difficulties when scientifically based dynamical models are used.

Objects in a Dynamical Spatial Environment

There are two major ways to view, and hence to model, the evolving spatial environment in which we live. The *object* view of the world sees individual objects located in a spatial domain and interacting through time with each other, often as a function of their distance apart. Thus, a household and its characteristics make up a unit of interest to census enumerators. This micro-datum is typically unavailable to social scientists, for confidentiality reasons. Consequently, the census data that are released are typically the *number of objects* in small areas, but not their locations. That is, a set of count data from small areas is released, which is simply an aggregated version of the object view of the world. The geographical extent (i.e., spatial support) of a small area can be stored in a Geographical Information System (GIS) as a polygon, and hence the spatial relationships between small areas and their associated counts are preserved in a GIS. [A GIS is a suite of hardware and software tools that feature linked georeferencing in its database management and in its visualization; e.g., Burrough and McDonnell (1998).]

Alternatively, the *field* view of the world loses sight of the objects and potentially has a (multivariate) datum at every spatial location in the domain of interest. Building on the census-enumeration example discussed above, we can define a *field* as the object density, in units of number per unit area, at any location. This is purely a mathematical construct because, at a given location, either there is an object present or there is not. Such a density can be estimated from a moving window, such that at any location the estimated density is the number of objects per unit area in the window at that location.

A useful way to think of the object view versus the field view is to imagine yourself in a helicopter taking off from a clearing in a field of corn. As the helicopter ascends, at some point it is no longer interesting to think of objects (e.g., corn plants), but rather to think of a field, literally and statistically (e.g., in units of bushels per acre). Then the temporal aspect is captured through the field's dynamical evolution during the growing season.

Sometimes the field view is the result of an aggregation of the object view, such as for population-density data. Other times, the field view is all that is of interest, such as for rainfall data where there is typically no interest in the individual raindrops. Again, a GIS is a convenient way to store data for a field, along with the spatial support to which a datum refers. Most of the exposition in this book (with the exception of Sections 4.2, 4.3, 4.4, and Section 6.6) is based on the field view. In general, spatio-temporal data may consist of measurements of both the field type and the object type. Modeling these data with coherent, spatio-temporal, random processes is the next frontier.

Uncertainty and the Role of Conditional Probabilities

The era of building (marginal) probability models directly for the data is coming to a close. A model of this sort defines a likelihood, from which inference on unknown parameters can be made. However, the likelihood does not directly recognize that data are a noisy, incomplete version of the scientific process of interest (see Section 2.1). This can be resolved by building a *conditional probability* model for the data, given the process, and then a separate probability model for the (hidden) process itself. From this perspective, it is clear that the likelihood is based on a *marginal probability* model of the data, where the scientific mechanisms are partially hidden by integration.

A lot of ink has been devoted to whether frequentist or Bayesian probability models are better. We believe that the bigger issue is whether marginal-probability or conditional-probability models should be used, and we are decidedly in the conditional-probability camp. As Statistics has become more a Science than a branch of Mathematics, conditional-probability modeling has shown its power to express uncertainties in all aspects of a scientific investigation. Such models have been called *hierarchical statistical models* (sometimes referred to as latent models or multilevel models); see Section 2.1.

Bayes' Theorem is a fundamental result in probability theory that allows an inverse calculation of the conditional probability of the unknowns (process and parameters) given the data (Bayes, 1763). Inference on the unknowns is based on this conditional probability distribution (called the *posterior distribution*), but the formula depends on a normalizing constant that is typically intractible (see Section 2.1).

Breakthroughs in the last 20 years have shown how an analytical derivation of the normalizing constant can be avoided by a judicious use of, for example, a Monte Carlo sampler from a Markov chain whose stationary distribution is the posterior distribution (see Section 2.3). This has made feasible the statistical analysis of scientific problems in the presence of uncertainty, based on hierarchical statistical models that can be of great complexity. But this comes with great responsibility; just because we can handle a lot of complexity, it does not mean that we should. This echoes our earlier comments at the beginning of this chapter, when discussing the Goldilocks Principle of model building.

Hierarchical Statistical Modeling

Hierarchical statistical modeling represents a way to express uncertainties through well defined levels of conditional probabilities. We follow Berliner's (1996) terminology: At the top level is the *data model*, which expresses the distribution of the data given a hidden process. This hidden process can be thought of as the "true process," uncorrupted by any measurement of it. At the level directly underneath the data model is the *process model,* which models scientific uncertainty in the hidden ("true") process through a probability distribution of the phenomenon of interest. It is quite possible that the process model is itself made up of submodels whose uncertainties are also expressed at sublevels through conditional probabilities. In a sense, the whole approach is a sort of analysis-of-variance decomposition that is more general than the usual additive decomposition given in standard textbooks (e.g., Scheffé, 1959). The result is a *hierarchical model* (HM); see Section 2.1.

The components of a HM are *conditional probability distributions* that, when multiplied together, yield the joint probability distribution of all quantities in the model. The quantities in which we are interested could be as simple as random variables and as complicated as space–time stochastic processes of random sets.

Of course, all the conditional probability distributions specified in the HM typically depend on unknown parameters. If a lower level (underneath the data model and the process model) is established by specifying the joint probability distribution of all the unknown parameters, then the HM qualifies to be called a *Bayesian Hierarachical Model* (BHM). This probability model at the lowest level, which we call the *parameter model*, completes the sequence: data model (top level) followed by process model (second level) followed by parameter model (bottom level); see Section 2.1.1. An alternative approach to specifying the parameter model is to *estimate* the parameters using the data. This might be called an *Empirical Hierarchical Model* (EHM), although historically it has often been called an empirical-Bayesian model; see Section 2.1.2. We prefer the nomenclature EHM, to contrast it with BHM.

Uncertainty and the Role of Statistics, Revisited

It is worth reflecting on how far we have come in this discussion of statistical modeling. We are not rejecting R. A. Fisher's paradigm of controlled scientific experiments; on the contrary, such experiments allow the statistician and scientist to build the suite of conditional probability models needed for hierarchical statistical modeling. For example, when water quality is measured through chemical assays, it is common to send duplicates and "blanks" (i.e., pure water) through the laboratory to gain knowledge about the measurement error in the data model. Furthermore, periodic recalibration of instruments guards against instrument drift over time and possible bias in the measurement errors. In agriculture, uniformity trials (where crops are grown but no treatments are applied to those crops) enable the scientist to build realistic, often spatial, process models. The HM paradigm enables a coherent use of all data and, using models of spatio-temporal statistical dependence, allows inference on parts where there are *no* data at all! Scientific relationships incorporated into the process and parameter models can mitigate the paucity of data. Furthermore, there is a self-correcting mechanism in hierarchical statistical modeling; when there is little known about the scientific relationships or there are poor-quality or few data available, then inferences have very low precision. That is, a signal in the process may be there, but if scientific knowledge or the data are limited, the HM approach will not let us discover it.

Looking at this from another angle, the best scientists collect the best data to build the best (conditional-probability) models to make the most precise inferences in the shortest amount of time. In reality, compromises at every stage may be needed, and we could add that the best scientists make the best compromises!

We conclude by saying that Science cannot be done "by the numbers." Good scientists require just as much inspiration as good artists, and indeed there is a view that they are symbiotic (Shlain, 1991; Osserman, 1995). To this we add Statistics, and particularly hierarchical statistical modeling, where data, Science, and uncertainty join forces.

Summary of the Book

This is a four-color book where not only is color used in the figures, but it is also used strategically in the text. Where appropriate, the data model is in green, the process model is in blue, the parameter model is in purple, and the posterior distribution is in red. Chapter 1 has introduced the broad philosophy of Statistics (with a capital "S") and its role in the scientific method. This is formalized in Chapter 2, where more notation, more methodology, and more statistical concepts are introduced. Readers have a choice at this point. Those unfamiliar with Statistics for temporal data could read Chapter 3, which reviews the fundamentals of temporal processes (i.e., dynamical systems and time series). Those unfamiliar with Statistics for spatial data could read Chapter 4, which reviews the fundamentals of spatial random processes. Those who are familiar with both could "Pass Go" and proceed to Chapter 5.

Chapter 5 introduces spatio-temporal statistical methodology through data, recognizing its roots in Science. Chapter 6 reviews the statistical models that have been used for analyzing spatio-temporal data. This book features dynamical spatio-temporal statistical models (DSTMs), and Chapter 7 gives a comprehensive exposition of them in the context of hierarchical statistical models. Implementation and inference for DSTMs in the hierarchical-modeling framework are presented in Chapter 8. Finally, a number of examples that illustrate Statistics for spatio-temporal data are given in the sections of Chapter 9.