METHODS: QUANTITATIVE METHODOLOGIES

Michael F. Goodchild, University of California, Santa Barbara

Michael F. Goodchild
Department of Geography
University of California
Santa Barbara, CA 93106-4060
USA

METHODS: QUANTITATIVE METHODOLOGIES

Keywords: quantitative revolution, Central Place Theory, deductive reasoning, inductive reasoning, statistical inference, spatial interaction model, optimization, statistical software

Glossary

Central Place Theory: a body of theory about the locations, sizes, and offerings of settlements in agricultural landscapes

data mining: the application of computational methods to large volumes of data in an effort to detect pattern

data model: a template or format for data

geographic information system: software for the analysis of geographic data

Google Earth: a web-based service for displaying information about the surface of the Earth

interval data: data values that can be subtracted to establish differences

least squares: a principle used to fit a mathematical function to data

linear regression: the fitting of a linear relationship between two variables in a sample

neural network: a form of analysis originating in artificial intelligence

nominal data: data values that serve only to differentiate

null hypothesis: a statistical proposition concerning a sample's relationship to its parent population

ordinal data: data values that establish order or rank

ratio data: data values that can be divided to establish ratios

self-organizing map: a form of analysis originating in artificial intelligence

social physics: the application of concepts from physics to social systems

spatial interaction model: a model of the interaction between an origin and a destination that includes the impeding effect of intervening distance

Tobler's First Law: the assertion that "all things are related but nearby things are more related than distant things"

Synopsis

The quantitative revolution of the 1960s stimulated interest in quantitative methods as tools of scientific investigation. The distinction between quantitative and qualitative methods is technical, but has come to signal a much deeper split in the methodology of human geography. Quantitative methods are indispensable tools for mediating the interaction between theory and experiment, within a scientific paradigm that emphasizes replicability and common understanding of terms. Statistical inference allows investigators to reason about the general properties of populations from evidence based on samples, but has significant difficulties when applied to geographic data. Much quantitative analysis is concerned with the fitting of mathematical functions to relationships, while the search for pattern and anomaly is increasingly viable in today's computing environments. Normative approaches that attempt to optimize some appropriate design function are popular, and the distinction between them and more

conventional positive approaches is often blurred. Software is now an essential part of quantitative methodologies.

Introduction

Ever since the development of counting and measurement systems, humans have recognized that quantities are more useful in many respects than qualities. The administration of land in the Fertile Crescent and the Nile Delta required measurements of length and area, and basic principles of geometry. Navigation required the ability to measure direction and position on the Earth's surface, and led to the science of map projections. It is clearly important for a retailer to be able to predict how many shoppers will patronize a store, and how much they will spend. The knowledge that an earthquake will strike is of far less value than the knowledge of precisely where and when.

In human geography, however, there is far less agreement on the value of quantification, and many of these issues are far from settled. Concern is often raised over whether humans can ever be predictable, and whether measurement is not of its very nature inhumane. Measurements of humans have been used for purposes that are now discredited, such as the 19th Century eugenics of Galton. Thus the urge to quantify has waxed and waned in human geography over the years. Social physics, or the adaptation of principles from physics to society, has occasionally caught the imagination, as it did in the 1950s with the work of Zipf, Stewart, Warntz, and others, who sought to apply Newton's Law of Gravitation and related principles to social phenomena. One of the more successful threads of this research led eventually to the models of spatial interaction that today provide invaluable predictions of traffic counts, retail store patronage, telephone traffic, and many other phenomena.

The heyday of quantification occurred in the 1960s during geography's Quantitative Revolution, when a generation of young scholars, many trained at the University of Washington, argued for a focus on theory and its empirical verification using quantitative statistical techniques. The spirit of the time is best captured in Bunge's *Theoretical Geography* and Harvey's *Explanation in Geography*. Quantification has made enormous strides since then, and today quantitative methods and scientific reasoning are as fundamental to human geography as they have always been to physical geography.

While the distinguishing characteristics of quantitative methodologies – a reliance on quantities, in the form of counts or measurements – may seem straightforward, on closer examination it is possible to distinguish several somewhat independent threads. The following sections discuss each of them in turn.

Qualitative and quantitative

Theories of data identify four basic forms: nominal, ordinal, interval, and ratio (directional data is sometimes identified as a fifth). Data are said to be nominal if their purpose is solely to distinguish one instance from another. Thus names of people are nominal, as are classes of land use. Nominal data may be numeric, telephone numbers and social security numbers being obvious examples, but in such cases it makes no sense

to perform arithmetic operations. The telephone number 8938049 is not more or less than the number 8933146. Data are said to be ordinal if their purpose is to express order, or to rank. Thus first may be better or more than second, and Class 1 land may be better than Class 2 land. Again, ordinal data may or may not be numeric: the colors of the rainbow provide an order, for example. Interval data are numeric, and differences between them are meaningful. Data expressed using the Celsius scale of temperature have interval properties, for example, and it makes sense to perform such arithmetic operations as averaging. Finally, both differences and ratios are meaningful for ratio data. It makes sense, for example, to say that a weight of 20 kg is twice a weight of 10kg.

In this framework nominal and ordinal data are qualitative, and interval and ratio data are quantitative. This distinction is precise, but it hardly explains the very strong cleavage that exists between proponents of quantitative and qualitative geographic methodologies, and the tendency for researchers to align themselves with one side or the other at an early career stage. After all, it is easy to make quantitative measurements of the area occupied by Class 1 land, an ordinal and therefore qualitative property, or to develop quantitative models that attempt to predict such qualitative properties as whether or not an individual will migrate in a given time period.

Instead the labels have come to characterize methodological differences that are far more profound and nuanced than the grouping of four data types into two categories would suggest. First, quantitative methodologies are strongly associated with the sciences. There is a willingness to believe in the process of statistical inference, which attempts to extract truths of general applicability from the study of limited samples, using rigorous reasoning. There is a belief in replicability, that investigators should reach the same conclusions from the same experiments; and in the need for universally accepted terms with rigorous definitions. In these respects quantitative human geography resembles physical geography and other environmental, life, and physical sciences in its fundamental assumptions and goals. Qualitative methodologists on the other hand may question many of these assumptions, and employ a wider range of techniques that stretch them to varying degrees. They may express doubts about the value of conventional philosophies of science when applied to social phenomena, and be willing to embrace knowledge that is not necessarily replicable; and will likely view their methods as uniquely applicable to the social sciences.

The role of theory

Quantitative approaches are firmly grounded in the notion of theory, or the possibility of general propositions about the domain of human geography. Much of the fuel for the quantitative revolution derived from Central Place Theory, a body of propositions developed by Christaller and Lösch, and addressing the patterns of settlements that could be expected to develop on an agricultural plain. Under fairly stringent assumptions about the behavior of consumers and entrepreneurs, Christaller and Lösch concluded that settlements should form a discontinuous hierarchy, with the smallest centers offering only limited ranges of goods and the most expensive goods being offered only in the largest centers; and that settlements should position themselves precisely at the nodes of a hexagonal network.

In the best scientific tradition, efforts were made to find confirmation of the theory by comparing its predictions to actual patterns of settlements, using areas such as Iowa that could be assumed to approximate a uniform agricultural plain. Efforts were also made to adjust the theory to specific circumstances, when for example the distribution of rural consumers was not uniform, due perhaps to spatial variation in agricultural productivity and the value of crops. Today, geographers think of Central Place Theory as a set of ideas that can be used to structure investigations of settlement patterns; but have largely rejected the notion that it might provide a precise model of the social world. In that sense the paradigm has come to resemble the one dominant in economics: propositions about the social world that are general but not necessarily in agreement with reality, but that nevertheless form a framework for understanding. Thus quantitative human geographers still believe in an interplay between theory and empiricism, in the best traditions of experimental science. The process of induction draws on exploration of the social world, suggesting general principles that accumulate into a body of testable theory; and the process of deduction draws on theorizing to suggest new principles that can be tested against reality.

If theory offers precise predictions, then it follows that quantitative methods will be needed to test them. Theory is necessarily general, so the methods used to test theory must involve large numbers of samples, and formal investigations of whether the samples confirm or deny the theory. Moreover, unlike theories about the physical world, it seems inevitable that theories about the social world must be less than perfect in their predictions – that the goal of perfect prediction is fundamentally unachievable, if only because humans are free to contradict predictions about their own behavior. Predictions that are less than perfect require large numbers of samples to confirm them, unlike perfect predictions which a single counter-example can refute. Thus there are many reasons why a human geography that is concerned with the discovery of general, testable truths should align itself with quantitative methodologies.

Statistical inference

Statistical inference originated in the life and physical sciences, and in concern over what could be concluded from experiments involving limited numbers of samples. For example, a field might be sown with two types of seeds, using 100 seeds of each type, and while one type appears to result in larger plants, there is overlap in the results – some plants from the better seed are smaller than some plants from the poorer seed. Is the apparent improvement a result of chance, or does it indicate a real superiority of one seed type over the other? A counter or null hypothesis is posed, in this case that the two seed types are equally productive, and the probability determined that the actual experimental result could occur if the null hypothesis is true. If this probability falls below some threshold, typically 5%, the effect is said to be statistically significant and the null hypothesis is rejected.

Several broad classes of inferential tests can be identified, depending on the nature of the null hypothesis. This example is a two-sample test, for which the null hypothesis is always that both samples were drawn from the same population. In other instances a

single sample is compared to some theoretical proposition, such as a population with a specific mean value, and the null hypothesis proposes that the sample was drawn from the theorized population. Statistical inference can also be used to evaluate a numerical relationship between two variables, in which case the null hypothesis proposes that the sample was drawn from a population in which the variables are statistically independent. Of particular interest to geographers are tests of spatial pattern against null hypotheses of spatial randomness.

The practice of statistical inference always carries risk. If a null hypothesis is rejected at the 5% level of significance then the researcher accepts a 5% chance of being wrong, in other words that the null hypothesis is in fact false. On the other hand if it is accepted, it is possible that a weak effect is nevertheless present, and a larger sample would show the opposite result. These two outcomes are described as Type I and Type II statistical errors, respectively.

Since its inception in the 19th Century, the practice of statistical inference has grown to become an almost essential part of any empirical research. In human geography most experiments are not controlled, as in the planting of two types of seed, but natural, in the sense that the conditions of the experiment are outside the researcher's direct control. Thus the two types of seed might correspond to two study areas, perhaps two cities or two ethnic groups. Under such circumstances the assumption that each sample is randomly and independently chosen tends to be untenable, particularly if samples are taken from locations close together in space.

Moreover, statistical tests in human geography are often plagued with the problems associated with drawing multiple inferences from the same sample. Consider a pattern of points denoting instances of a disease, and suppose that the researcher wishes to determine if clusters occur – if areas that appear to have higher density do indeed have anomalous properties. It is possible to test any area against a null hypothesis of uniform density, but how should the researcher decide which areas to test? Inevitably the choices will be determined by the distribution of the very phenomenon one is proposing to test – in other words, the null hypothesis is untenable a priori.

Reference has already been made to the fundamental complexity of human behavior, and the impossibility of perfect prediction. In the world of human geography one suspects that virtually any effect is detectable given enough data – in statistical terms, that with enough data one can almost always refute a null hypothesis, and that failure to reject is almost always a Type II statistical error. Moreover the acceptance or rejection of a null hypothesis confounds the strength of any effect with the size of the sample. Nevertheless establishing the significance of an effect has become a primary goal in much of the quantitative literature. Peter Gould's 1970 paper "Is *statistix inferens* the geographical name for a wild goose?" is a compelling review of these arguments for and against statistical inference.

Statistical models

Another thread of the quantitative revolution concerned the ability to infer generalities from pattern, through the analysis of patterns of features on the Earth's surface. Certain general principles, it was argued, would leave characteristic footprints, such as the hexagonal network of settlements of Central Place Theory; and analysis of pattern could therefore lead to inference about theory. But given the complexity of the human world, any such patterns could be expected to show erratic distortions, which might be modeled using statistical principles. Accordingly, much energy was expended in the 1960s in efforts to fit statistical models to the distributions of settlements, in order to show that while perfect hexagons did not exist, some degree of hexagonality or other kind of order could still be detected. Geographers were able to show that the numbers of settlements in each cell of a grid laid over Iowa were indicative of a pattern that was not random, but tended towards a uniform spacing of settlements; and similar methods were used to detect clustering in patterns of disease, ethnicity, and many other phenomena.

In such cases, however, the set of statistical models against which real patterns can be compared is extremely limited. More recently geographers have become interested in various kinds of simulation models that exploit the power of computers to represent complex behaviors and interactions, and to compute the patterns that they produce on the geographic landscape. In effect, such simulation models replace the simplistic null hypotheses of traditional statistical analysis with more realistic hypotheses, and success results when a simulation matches observation and the hypotheses are accepted, in contrast to the double-negative statistical tradition of rejecting a null.

Geographers are particularly interested in the kinds of patterns that result when values are assigned to reporting zones such as counties or census tracts. For example, such patterns arise in the distribution of income or population density, and several hundred statistics are produced for every reporting zone in the aftermath of every decennial census. The characteristic known as positive spatial dependence – a tendency for nearby areas to have similar values – is almost universally observed in such data, and has led to the assertion known as Tobler's First Law. Specialized statistical tests for the presence of spatial dependence have been developed, and are widely used by geographers, based on rejection of a null hypothesis of independence.

Curve fitting

Consider a quantitative relationship of the form $y = f(\mathbf{x})$ where y is some dependent variable, f is a function, and $\mathbf{x}$ is one or more independent variables. Such a relationship might describe the effects of income x on economic deprivation y (in this case f would be expected to be inverse, in other words an increase in x would result in a decrease in y). In another example y might represent average income in a county, and x years of education. In a classic study of this nature, Openshaw and Taylor examined the relationship between percent Republican voters and percent 65 and over in the 99 counties of Iowa. The model $y = f(\mathbf{x})$ might include spatial or temporal lags, if it is hypothesized that y depends on levels of education in previous time periods, or in adjacent counties.

The simplest possible relationship between two variables is linear, and in the absence of other knowledge or hypotheses a researcher might evaluate the relationship $y = a + bx$ where a and b are constants. Linear regression is a time-honored technique for determining a and b from real data, in other words of fitting the curve f. It proceeds by a process known as least squares, by choosing a and b to minimize the sum of squared deviations between observed values of y and their values predicted from the model. Linear regression has spawned a host of related techniques for handling more complex models, more specific concepts regarding deviations from the model, and more complex hierarchical structures. Of particular interest to human geographers are versions of the model that incorporate spatial lags.

Figure 1 shows the relationship between median value of housing and percent black by county from the 1990 U.S. census. Figure 1a shows the counties of California, and indicates a clear quantitative trend, counties with higher percent black having higher median value. Figure 1b shows the counties of Alabama, and indicates the opposite trend. In California the counties with the highest percent black are urban and comparatively wealthy, whereas in Alabama the counties with the highest percent black are rural and comparatively poor.

[Figure 1 about here]

Curve fitting is invaluable in prediction, in making estimates of the value of some outcome y based on scenarios involving one or more inputs $\mathbf{x}$. One of the more successful applications in human geography has been to the spatial interaction model, which attempts to predict interactions or flows between an origin and a destination based on properties of both and of the trip or separation between them. The classic spatial interation model has the form $I_{ij} = E_i A_j f(D_{ij})$ where I_{ij} denotes the interaction between origin area i and destination area j, E_i is a factor characterizing the origin area's propensity to generate interaction (its emissivity), A_j is a factor characterizing the destination area's propensity to attract interaction (its attractiveness), and d_{ij} is a measure of the separation of i and j. The function f will be a decreasing function of distance or separation corresponding to the impedance associated with geographic separation.

The spatial interaction model has been applied to a host of phenomena from journeys to work to migration and social interaction. Data from these areas have been used to calibrate the model, in other words to determine the values of various unknowns, such as the form of the impedance function f, by comparing the model's predictions to real data. Various elaborations of the model have been described, and there are several theoretical justifications for its functional form. It is routinely used to predict shopping behavior under scenarios involving new development, closure, and modifications to the transportation network.

Number crunching

One of the most fruitful areas for development of quantitative methodologies in recent years has been in the area somewhat pejoratively described as number crunching – the mining of large amounts of data in search of pattern and ultimately hypotheses, in a

decidedly inductive spirit that is largely independent of theory. One of the strongest proponents of this paradigm in human geography is Openshaw, whose series of Geographical Analysis Machines are designed to submit data to large numbers of exploratory hypotheses, many of which may not make any immediate sense in any recognized theoretical framework.

An early version of this approach was popular in the 1970s, particularly at the University of Chicago, a traditional center of quantitative social science. Factor analysis was devised in the 1930s as a tool for examining large matrices of data in a search for fundamental but hidden dimensions. For example, one might submit the results of a large number of psychological tests to such an analysis in an attempt to identify what one might claim to be the underlying dimensions of personality. A similar approach was adopted in examining large amounts of census data in an effort to discover the underlying dimensions of geographic variation in human society. Studies were conducted on many cities, and two dimensions consistently emerged: a composite of various indicators of wealth, and another of indicators of life-cycle stage. Critics pointed to the unknown effects of the variables chosen by the census for tabulation, the unknown effects of the reporting zone boundaries, and the arbitrarily linear nature of the analysis.

With the phenomenal growth of computing power and data availability over the past two decades, such inductive methods have become increasingly popular. Neural networks began as an effort to provide a crude model of how the brain might operate, but have been adopted as theory-neutral tools for the analysis of large data sets, with some success in the general area of prediction. Self-organizing maps are another product of research in artificial intelligence that have appealed to geographers as methods for discovering pattern, and perhaps hypotheses, in large data sets. The term data mining has been popularized in this context.

Optimization

The quantitative revolution's interest in Central Place Theory stemmed largely from its potential as an explanation of settlement patterns – of why settlements appeared on agricultural landscapes in the observed locations and sizes, and offering particular combinations of goods. From time to time, the same theory has been used for a quite different purpose, as a basis for planning new landscapes, when decisions on locations, sizes, and perhaps offerings of goods are in the hands of planners. For example, planners were required to make decisions about the locations of settlements during the draining of the Dutch polders in the mid 20th Century.

Similar concern for design has underlain many other applications of quantitative methods in geography over the past half century. Geographers have contributed to the literature on the optimal location of linear facilities such as highways, pipelines, and power lines; point facilities such as schools, fire stations, and retail stores; and area facilities such as nature preserves and voting districts. Many of the methods fall under the general heading of operations research, a subdiscipline that is also found in transportation, industrial engineering, and management science.

The use of optimization for design purposes has been termed normative geography, to distinguish it from the more traditional interaction between theory and experiment that one might term positive geography (in line with the prevailing positivist philosophy of science). But the boundary between these two paradigms is often blurred. Consider, for example, an application of Central Place Theory to design. While the planner might determine the locations and sizes of settlements, the success or failure of an offering of a particular good at a particular settlement will be determined in the real world, as will the choices consumers make about which settlements to visit and how much to spend. But successful planning clearly requires that these aspects be predictable, even though they are not controlled. Thus positive geography must be invoked to provide the predictions, in order to ensure the success of the normative plan.

Data modeling and software

The early quantitative geographers worked with pen and paper, aided from time to time by tables of logarithms and slide rules. But the advent of powerful computing machines, which became available on university campuses in the early 1960s, undoubtedly added momentum, and today computing and quantitative methods are inseparable. Gone are the days when instructors could insist that students perform tests by hand, to ensure that they understood the mechanics, before being unleased on computing environments. Today, statistical software is universally available, and geographers are aided by such specialized applications as geographic information systems (GIS), software for exploratory spatial data analysis, and web-based services such as Google Earth.

Within the world of statistical computing the dominant data model is the table. Cases, observations, or samples are arrayed in the rows, and the columns hold the various measures and counts associated with each row. Tables are the primary format of data distribution through such institutions as the US's Interuniversity Consortium for Political and Social Research (ICPSR) or the UK's Essex Data Archive, and are the primary data model for statistical software.

Although tables are a powerful mode of representation, they are unsatisfactory in several respects for research in human geography. If the data refer to points, lines, or areas, then it is likely that the locations and geometric forms of the individual cases will be captured, as an aid both to cartographic display and to analysis. While the locations of points can be represented as pairs of coordinates, lines and areas require more complicated structures that do not fit neatly into tables. Handling such complex geographic features has become the domain of GIS, and a common format known as the shapefile has become the de facto standard, despite its origins in the products of a single software vendor.

Further reading

Bunge, W. (1966) *Theoretical geography*. Second Edition. Lund: Gleerup.

Christaller, W. (1966) *Central Places in Southern Germany*. Translated by C. W. Baskin. Englewood Cliffs: Prentice Hall.

Clark, W. A. V. and Hosking, P. L. (1986) *Statistical Methods for Geographers*. New York: Wiley.

Fotheringham, A. S., Brunsdon, C., and Charlton, M. (2000) *Quantitative Geography: Perspectives on Spatial Data Analysis*. Thousand Oaks: Sage.

Ghosh, A. and Rushton, G., editors (1987) *Spatial Analysis and Location-Allocation Models*. New York: Van Nostrand Reinhold.

Gould, P. R. (1970) Is *statistix inferens* the geographical name for a wild goose? *Economic Geography* 46: 439-448.

Griffith, D. A. and Amrhein, C. G. (1991) *Statistical Analysis for Geographers*. Englewood Cliffs: Prentice Hall.

Haining, R. P. (2003) *Spatial Data Analysis: Theory and Practice*. Cambridge: Cambridge University Press.

Harvey, D. (1970) *Explanation in Geography*. New York: St Martin's Press.

Lösch, A. (1954) *The Economics of Location*. Translated by W. H. Woglom and W. F. Stopler. New Haven: Yale University Press.

Maguire, D. J., Batty, M., and Goodchild, M. F. (2005) *GIS, Spatial Analysis, and Modeling*. Redlands: ESRI Press.

Miller, H. J. and Han J., editors (2001) *Geographic Data Mining and Knowledge Discovery*. New York: Taylor and Francis.

O'Sullivan, D. and Unwin, D. J. (2002) *Geographic Information Analysis*. New York: Wiley.

Figure captions

1. Scatterplots of the counties of (a) California and (b) Alabama, showing quantitative relationships between median value of housing and percent black. The dashed lines represent fitted linear trends.

2. Two realizations of a model of uncertainty in the four surveyed corner points of a simple land parcel.