8. Spatial modelling and forecasting

Subhrajit Guhathakurta, Ge Zhang and Bon Woo Koo

1. INTRODUCTION

Spatial modelling is the process through which spatial data are analyzed to derive patterns and relationships among spatial objects. These spatial objects can be fixed, such as land, buildings, rivers and roads, or dynamic, as in the movement of people and vehicles in space. Spatial modelling has wide applications across many disciplines including transportation, epidemiology, ecology, archaeology, and urban and regional planning, which is the focus of this chapter. Although the techniques of spatial analysis have particular nuances when applied to various disciplinary problems, they share a common foundation, built upon a range of statistical methods. The objective in this chapter is to show how spatial analysis is used in problems that commonly concern urban planners who are seeking to understand how the future will unfold under various scenarios. The focus is on developing a framework for understanding the current context of spatial analysis and outlining several undercurrents in theory and methods that have far-reaching implications for spatial modelling and forecasting in planning.

It is important to situate the discussion of spatial modelling and forecasting within the current ethos of planning in the public domain. There are reasons to conclude that planning is at a crossroads; so are many other academic disciplines. The evidence is all around us. In the past decade, we have witnessed growing social and political conflicts across the globe, the diminishing effectiveness of local, regional and global institutions leading to a lack of trust in their abilities, the continuing debates over climate change and the near paralysis of action in addressing it, and an unprecedented housing market-led financial-economic meltdown. Most importantly, our faith in science and human knowledge systems has been shaken. Urban planners especially are at the receiving end owing to this convergence of multiple crises given their role in charting a livable future. Spatial modelling and forecasting, particularly of the long-range variety, have been impacted understandably by the complex socio-political economic uncertainties that have buffeted this era.

Yet, more recently, we have also witnessed the maturing of several new tools, technologies and social processes that have all coalesced to offer exciting new ways of addressing urban problems. The big data revolution is gathering steam together with new methods for querying, filtering and analyzing data in near real time. Social media are transforming ways in which people communicate and obtain information, creating new ways of legitimizing facts and 'alternative' facts. While the increasing abundance of data has provided the impetus for open-data portals, many critical urban data-generators have exerted tight controls by privatizing information about public use of urban space. The sharing economy is expanding with new products and service exchanges coming online in quick succession. Another significant and highly consequential development is the impending introduction of commercial self-driving cars. Given past experience with new transportation technologies, there is little doubt that this new mode of transport will transform urban form, land use and social choices about travel and location behavior.

The legacy spatial models are losing their appeal, especially because long-range planning tools have not delivered the promise of a better future. Advances in data mining and data fusion technologies have led to opportunities for urban spatial modelling that can serve a range of communities. First, the open data/open tools movement has picked up momentum and many urban regions are sharing their data and models with other communities. This has created a community of experts that are also sharing stories about the nuances and successes of using the data and tools, which help smaller communities in their modelling efforts. Second, several important sources of spatial data have now become ubiquitous, that is, available everywhere for all areas of the globe. These data include textual data from social and print media, satellite images including multi-spectral and hyperspectral data, Google Street View, mobile phone usage, volunteered geographic information (VGI), data from local sensors and video feeds, and other publicly available archived data. Planners have been slow to adopt the new sources in their spatial modelling efforts, but that is changing. Finally, the idea of making cities 'smart' has enticed private firms and city officials to focus on solving urban problems with the help of data and technology. The smart city efforts have been inclusive thus far, given that smaller communities are more numerous and offer a large market for firms engaged in delivering smart technologies. The question now is what planners can do to leverage this enthusiasm and shape it in a way for enabling just, sustainable and livable communities. This chapter engages in that conversation too.

2. URBAN SPATIAL MODELS AND POST-NORMAL SCIENCE

Planning needs to respond to a series of unprecedented challenges related to both mitigating and adapting proactively to uncertain futures that are expected to be qualitatively different from those currently. This is complicated by the more significant issues impacting cities and regions, such as climate change, global financial crises or terrorism, being forces outside their borders. In addition, the potential impacts of these disruptions are so significant that localities or even states cannot generate the resources locally to address them adequately. Also, there are already many unmet needs and relatively few resources in most localities. The question of what priority should long-range planning for sustainable development receive and how that planning actually occurs is becoming the wickedest of modern problems (Rittel and Webber 1973).

A wicked problem, that is, a problem in which the definition of the problem is contested, and solutions are unknown or untested, was reflected in the debates around the epistemological condition known as post-normal science (PNS) (Funtowicz and Ravetz 2003). Thomas Kuhn, in his well-known treatise on the structure of scientific revolutions, described scientific advances as a cyclical process that iterates through normal and post-normal periods. According to this theory, in the normal state, science is managed through incremental processes that build on an agreed upon epistemological foundation. Uncertainties and value propositions are managed adequately within this framework. The post-normal state, in contrast, is characterized by conditions in which 'facts are uncertain, values in dispute, stakes high, and decisions urgent' (Funtowicz and Ravetz 2003, p. 349).

The scientific facts in post-normal periods are legitimized through extended peer communities of stakeholders with different perspectives and interests who coalesce around the problem being addressed. The solutions are grounded, that is, they are specific to time, place and constituencies, and they cannot be generalized in the way normal scientific laws are formulated. Making progress in resolving the critical (and often wicked) problems requires new methodological approaches. In these approaches, uncertainties are accepted and managed while values are made explicit. The solutions are not just scientific deductions but an interactive coalescing of interests. Action based on the knowledge derived from this process is legitimized not just by the peer scientific community but by the extended peer community that includes scientists, policy-makers and other stakeholders.

Spatial modelling that is primarily designed to produce future scenarios by adjusting certain empirically derived parameters is acutely affected by the post-normal ethos of this era. In the urban planning domain, spatial models have been the primary tools for forecasting land-use transportation scenarios. The models have a long and checkered history that started in the 1960s as gravity-based spatial interaction models (Goldner 1971; Lowry 1964; Putman 1983; Putman and Chan 2001). More recently, these models have developed along different methodological directions including: (1) site suitability, such as What If? and CUFS-I (Klosterman 1999, 2008; Landis 1994); (2) cellular automata, such as SLEUTH (Chaudhuri and Clarke 2013; Silva and Clarke 2002); and (3) disaggregated random utility based econometric models such as UrbanSim (Waddell 2002) and the Production Exchange Consumption Allocation System (PECAS) modelling environment (Hunt and Abraham 2005). The data requirements and the ease of calibration of each of these approaches are based on the spatial granularity, the extent of the region being modeled, and the number of factors considered as drivers of change for future land-use and transportation scenarios.

While a great deal has been written about the difficulty in calibrating spatial land-use transportation models owing to stringent data requirements and the need to have detailed local knowledge for implementing these models, far less attention has been directed to the question of what scenarios are useful to consider in particular contexts. Spatial models are mathematically or stochastically based projections of a future state (Godet and Roubelat 1996; Houghton et al. 2001). They only generate the future state of scenarios that are first contemplated by the model developers together with experts and stakeholders (Maack 2001). Scenarios are often designed to challenge current assumptions and world views and imagine possible outcomes that span the range from likely to highly unlikely (but plausible). Spatial models enable assessment of each scenario to determine those that are desirable and to identify decisions that could lead to those desirable scenarios instead of the alternatives (McCarty et al. 2001; Santelmann et al. 2001; Steinitz et al. 2003).

Given that spatial models are only as useful as the number and characteristics of the scenarios they are designed to assess, this chapter proposes an approach for scenario development that leverages multiple perspectives and expectations about the future from stakeholders to generate many scenarios. The scenario development approach addresses potentially high-impact future events that have few parallels in everyday experience. This approach is also key to advancing decision-making in the post-normal era. The method for scenario development and its relationship to forecasting is discussed in section 5.

3. NEW TYPES OF SPATIAL DATA

Until recently, spatial data at a granular geographic scale have been expensive to obtain and process (Axinn and Ghimire 2011; Bifulco et al. 2014; Krieger et al. 1997; Leeuw and Collins 1997). However, recent advances in electronic sensors, image processing, and machine learning are advancing the acquisition, processing and analysis of spatial data (Curtis et al. 2013; Seo et al. 2008). For example, algorithms in geostatistics, such as Bayesian maximum entropy, have been applied for estimating future water usage based on trends and projections of population density (Lee et al. 2010). Similarly, image interpretation technology can generate a detailed analysis of land cover and land use for several square kilometers within a few minutes from satellite imagery (Karnieli et al. 2008; Vittek et al. 2014). General purpose machine-learning frameworks such as TensorFlow (https://www.tensorflow.org/) and computer vision-based techniques like PSPNet (https://arxiv.org/abs/1612.01105) and several others are being developed in quick succession offering new ways of characterizing the built environment from images like Google Street View (GSV). Information abstracted from social media platforms like Twitter also offer an excellent source of data from users (Guhathakurta et al. 2019).

In most of the machine-learning based techniques, the models need to be trained by feeding a large number of classified and labeled objects observed in the images. Once trained, they can be used to identify similar objects or pixels from all other images from the same source. The new techniques are mostly based on, and can be applied to, very large datasets – often referred to as big data. Despite its extensive use, the definition of big data has not reached a consensus. Nonetheless, the characteristics and sources unique to big data are relatively well identified. In the literature linking big data and cities, big data are often viewed from two frameworks: one that describes the characteristics of big data, and the other that describes the sources of big data (Kitchin 2014).

A popular way of distinguishing the characteristics of big data from conventional data are the three Vs – standing for volume, velocity and variety (De Mauro et al. 2015). Volume refers to the large size of data; 'social media, e-commerce, and the Internet of Things generate approximately 2.5 quintillions of bytes per day, an amount that equals 100 million Blu-ray discs, or almost 30,000 GB per second' (Bailly et al. 2018, p. 1524). Velocity refers to the frequency or interval at which new data feeds are generated and analyzed. This is made possible by the deployment of digital sensors and the Internet of things (including computers and smartphones), which can generate a large volume of data in real time, transmitted to servers through a wireless network. Variety refers to the diversity of data type or the structural heterogeneity in a dataset that encompasses text, image, video and audio, to name a few (Mohammadi et al. 2018). More recent studies add two more Vs, veracity and value, which refer to the fact that the quality of data is uncertain and that data can be valorized, respectively (Bailly et al. 2018).

As for the sources of big data, they can be grouped into three general categories: directed, automated and volunteered (Kitchin 2014). Directed data are 'generated by traditional forms of surveillance, wherein the gaze of the technology is focused on a person or place by a human operator' (Kitchin 2014, p. 4). They include remote sensing and LiDAR datasets that are generated by airplanes or satellites moving along the human-specified paths and collecting data at predetermined intervals. Google Street View images can also be classified as directed data. Automated data are 'generated as an inherent, automatic function of the device or system' (Kitchin 2014, p. 4). Examples of automated data include global positioning system (GPS) data transmitted from cell phones, transaction records and clickstream on e-commerce websites such as Amazon, and tap-in and tap-out records from public transportation systems. Volunteered data are, unlike the last two sources, generated by users. They include postings on social media such as Twitter, Facebook and Reddit, images and videos uploaded by users, reviews and tips left on Google Places and Foursquare, and user-contributed large databases such as OpenStreetMap and PlacePulse database.

Many of these big data are unstructured. Only about 5 percent of all existing data are structured (that is, tabular data in a spreadsheet or similar formats) while the rest is not in these formats (Cukier 2010; Gandomi and Haider 2015). Unstructured data, such as images, audio, video and unstructured texts, often need to be translated to into structured formats required by analysis and modelling conventions (Gandomi and Haider 2015). For example, although Google Street View images can be a valuable source of information, the raw images themselves are not useful in spatial modelling because they are not quantifiable and thus not interpretable by the models. Images become useful data after they are processed into data formats compatible with the models (for example, by counting the number of people in images or categorizing and labelling them according to research objectives). These processes often require machine learning or similar algorithms. Machine learning is not only an analysis technique but also an important part of the data management process, converting unstructured, unusable data into formats compatible with existing analysis and modelling techniques.

Many data coming from sensors and wireless networks (for example, smartphones) are inherently spatial and spatiotemporal (Jardak et al. 2014). A significant share of big data is comprised of geospatial data, and 'the size of such data is growing rapidly at least by 20% every year' (Lee and Kang 2015, p. 74). A study in 2012 noted that Google generates about 25 petabytes of data per day, and a significant portion of the data has spatiotemporal components (Vatsavai et al. 2012). In addition to relational data, the spatial dimension of big data offers important insights and allows researchers to gain greater value from the data by, for example, joining different datasets that are otherwise disconnected.

Geospatial data can be categorized into three forms: raster, graph and vector data (Shekhar 2012). Perhaps the most well-known type of raster geospatial data is remote-sensing data from aerial vehicles and satellites. In addition to traditional applications of these data, some novel applications are being tested, such as using satellite images and machine learning to predict socioeconomic measures, such as average household consumption (Jean et al. 2016). The most common example of graph data is road network that consists of nodes (for example, intersections) and edges (for example, street segments) (Lee and Kang 2015). Finally, vector data includes points, lines and polygons. In GIS, these data types form building blocks with which other big data sources are linked with one another. For example, geo-located tweets are often converted to point data which then can be joined with other geospatial data such as Google Street View images, GPS trajectories, road networks, city boundaries or remotely sensed data. With appropriate theoretical underpinnings, such as locational information, spatial proximity and the connections created by proximity provide a critical informational axis on which different sources of data that are otherwise disconnected can be joined together.

Although not all big data sources are geospatial, efforts have been made to add locational information to some non-geospatial data. For instance, while some tweets are geotagged and therefore can be spatially located, the proportion of such tweets is less than 0.6 percent (Lee et al., 2014). To add locational information to such non-geotagged Twitter data, researchers have used other geospatial big data sources, such as Foursquare, to build a model that can estimate the location from which Twitter posts are written (Lee et al. 2014). Similarly, images uploaded on various image-sharing platforms, such as Flickr, often do not contain latitude/longitude information. It has been shown that the location of images without locational data can be estimated using, for example, various visual (for example, contents of an image) and/or textual information (for example, image tags, comments added to an image, and tweets posted by an image uploader) (Cao et al. 2015; Hauff and Houben 2012).

In summary, directed, automated and volunteered big data can offer detailed spatial information about a large (volume), near-continuous (velocity), and diverse (variety) cross-section of urban experience. Much wider aspects of urban life have become measurable. The world generates over 6000 tweets per second, some of which can be automatically collected and processed/converted into formats for research needs (Mohammadi et al. 2018). Street view images can be used to train machines to predict people's political orientation (Glaeser et al. 2018), and how people perceive the given built environment (Naik et al. 2014).

These predictions can be treated as the end product or as input data to subsequent analysis or for spatial models. Also, the velocity of big data has a potential to alleviate the dependence on past events in predicting the future. As Dalton (1986) and Friedmann (1978) argued, the dependence on past events can be problematic because people can 'think, learn, and react emotionally', which can alter the patterns embedded in any data and invalidate the assumption needed to make robust predictions – that the patterns found in past data will continue into the future. Although the velocity of big data may not eliminate the concern, it may reduce the difference in patterns between data (that is, past events or measurements) and reality by narrowing their temporal gaps and allowing rapid updates to the previously extracted patterns.

It should be noted that big data by themselves are of little value. 'To be useful, data must be analyzed, interpreted, and acted on. Thus it is the algorithms – not data sets – that will prove transformative' (Obermeyer and Emanuel 2016, p. 1216). The next section discusses modelling techniques that can harness the potential that big data has.

4. MODELLING WITH SPATIAL DATA

Urban spatial models generate future states of a region by showing how its essential functions are going to change and how these functions may be accommodated spatially (Deal and Schunk 2004; Fang et al. 2005; Verburg and Veldkamp 2005; Waddell 2002). The projections of future states give us important information about where people live and work, how they travel, what buildings they use and how they spend their leisure. By projecting individual and social choices, behavior and activity locations, urban spatial models allow us to estimate the amount of energy, water and open space that will be needed. However, urban models only project what we can imagine. It is only our imagination and the resulting mental models that ultimately allow us to contemplate and envision urban sustainability. Spatial modelling is an important part of the planners' toolkit for making our mental models correspond to reality.

4.1 Conventional Spatial Modelling Techniques

A wide variety of spatial modelling techniques have been part of the standard toolkit bundled with most geographical information systems (GIS). This toolkit has been growing rapidly in the last two decades, especially since new statistical models to account for spatial dependence using spatial weights were developed, such as Getis Ord (G*), Moran's I and Local Indicators of Spatial Association (LISA) (Anselin 1995; Anselin and Rey 2014; Ord and Getis 1995). Spatially adjusted or geographically weighted regressions are commonly used in the analysis of geographically dispersed data. Some of the common conventional spatial modelling approaches are described below.

Overlay analysis

Overlay analysis is one of the most useful and simplest spatial modelling methods in planning (Collins et al. 2001; Engelen et al. 1997; Mosadeghi et al. 2015). This analysis is often referred to as the suitability approach, which originated from landscape analysis popularized by Ian McHarg's most celebrated book *Design with Nature* (McHarg and Mumford 1969). The most typical and common example of this method is land-use suitability analysis (Malczewski 2004). Here, the future land-use functions are predefined, based on empirical analysis or experience, such as land for biodiversity and habitat protection, land for agricultural activities and land for economic or residential development. For example, soils, climatic conditions and water availability can influence agricultural production. The data for each of the factors are collected, normalized, weighted and calculated as a composite score which can be shown on the map. The relationships between factors and the composite score are mostly linear. Each factor has a weight. The values of the factors are multiplied by the corresponding weights and then summed up to a composite score.

The sites or areas with the highest score are then selected as the best place for specific land-use development. Miller et al. (1998) used overlay analysis and land suitability analysis to identify the suitable place for greenway development. The basic tenets of the land suitability approach are implemented in the software package called What If? developed by Richard Klosterman (Asgary et al. 2007; Klosterman 2008). Klosterman emphasizes that What If? is not a forecasting tool but a planning support tool that shows 'what' would happen 'if': (1) particular development policies are enacted; (2) growth assumptions prove to be true; and (3) the user-supplied suitability scores are appropriate and reasonable. A variation of this approach was implemented in California Urban Futures (CUF-1) model (Landis 1994, 1995). Suitability for development in CUF-1 was based on profitability rather than physical and environmental constraints assumed by the user. Other researchers have used overlay analysis to identify the undevelopable areas based on hazards, such as earthquakes, typhoons, landslides and floods (Tsai and Chen 2011).

Cluster analysis

Cluster analysis is a spatial modelling method to find the hot spot or pattern of the urban components in the planning field (Huang et al. 2007; Romesburg 2004; Yu et al.

2014). The idea of cluster analysis is that the geographic units can be assigned into one group based on the spatial distance or the similarity in one or some of their attributes. For observing the location of these clusters, several local indicators have been suggested including the local Moran's I, which is the basis for calculating the LISA (Anselin 1995). The LISA has become a popular technique to determine hot and cold spots within a region for any spatially distributed phenomenon. There are, however, several other methods for determining concentrations of particular features or attributes across space, such as the Getis and Ord (1992) statistics, Gi and Gi*.

The attractiveness of the LISA, as explained by Anselin (1995), is that it decomposes Moran's I measure into the individual contributions to that measure by each observation. Therefore, the sum of the LISA scores overall for locations is proportional to the corresponding global statistic. The calculation of the LISA is easily accomplished with the help of several software packages including the freely available GeoDa package developed by Luc Anselin (https://spatial.uchicago.edu/software) and from the geostatistical tools in Esri's ArcMap.

Interpolation analysis

Interpolation analysis is used to predict spatially dispersed unknown data points based on the known sampling data (Anselin and Gallo 2006; Blanchet and Lehning 2010; Holdaway 1996). The common methods used in interpolation analysis are inverse distance weighted (IDW), rectangular, natural neighbors and kriging (Lu and Wong 2008; Rovatti et al. 1998; Sambridge et al. 1995; Stein 2012). The premise for each of these interpolation methods is that the unknown points spatially closer to known sampling points have smaller differences in their values than the unknown points farther away.

New techniques for interpolation analysis are also being developed. Hu et al. (2013) tested a multifractal inverse IDW method to predict land price. Multifractal interpolation preserves locally specific, high-frequency information, which is lost in any conventional moving average methods such as kriging and ordinary IDW (Cheng 1999). This method accounts for both spatial association and the scaling dependency from a multifractal point of view by assuming that the statistical behavior of a spatial variable changes as the measuring scale changes. Hu et al. (2013) suggest an effective alternative for predicting land prices compared with ordinary IDW and kriging methods.

Several interpolation techniques have been proposed to address unique challenges in different types and formats of spatial data (Mitas and Mitasova 1999; Wu and Murray 2005). Data that are spatially discontinuous and noisy are especially difficult to model accurately. In these cases, different methods can produce widely divergent results. Therefore, in-depth knowledge about the phenomenon being modeled is critical. Synthesis of data from various sources, together with the application of multivariate models, has resulted in better estimations of complex spatial variations.

Comprehensive models using multivariate regression techniques

Regression analysis estimates the relationship between one or more explanatory variables and the target-dependent variable comprising known sampling points (Huang et al. 2010; McMillen 2004; Oliveira et al. 2012). Once a mathematical function is estimated using regression, it can be used to predict the target variables for unknown points based on the known explanatory variables for those points. Most urban land-use forecasting models are based on the random utility theory popularized by Daniel McFadden, which won him the Nobel Prize in Economics. Models based on random utility are discrete choice models that are estimated with the help of multinomial logistic regressions. For example, Hu and Lo (2007) estimated urban growth transitions of the Atlanta metropolitan region using explanatory variables such as population density, distances to nearest urban clusters, activity centers and roads, high-density urban uses, distance to the CBD and the number of urban cells in a 7 × 7 cell window. Their results were represented in a land-use probability map which could be used to predict the future land use in Atlanta.

A disaggregated and spatially as well as temporally dynamic model that is becoming a gold standard in metropolitan planning organizations is UrbanSim (Waddell 2002). UrbanSim is capable of forecasting land use, households and jobs at the parcel, census block or zonal levels. It has been in active development since the late 1990s, first at the University of Washington, then at the University of California at Berkeley, and now, since 2016, at Urbansim Inc., a privately owned company. In 2005, UrbanSim was completely re-engineered to adopt a more extensible and modular platform called OPUS, based on the Python code. More recently, a cloud-based interactive platform called Urban Canvas was launched to vastly simplify the process of model estimation and visualization (see Chapter 27 in this volume). This platform includes pre-built models and data for almost all metropolitan areas in the USA, simplifying access to data and computing resources.

UrbanSim introduced several innovations in the field of urban environmental modelling. It consists of several interacting models that reflect the decisions of each household, business and developer. The interplay of these individual decisions leads to the final outcome of their location in space. UrbanSim explicitly models each household's decision to locate in or relocate to a parcel, census block or zone in the metropolitan area. Similarly, every business is identified, and location and relocation decisions estimated. New developments are predicted by modelling the change in land prices. The changes in travel behavior and the impact on locational and development decisions are explicitly incorporated in the modelling platform by interfacing UrbanSim with a travel demand model. Therefore, UrbanSim is usually run in tandem with a separate travel demand model given that this feature is not included in the current platform.

Several other models are also being developed and tested at different parts of the world. Some, such as TLUMIP2 (Oregon, USA), IRPUD (Dortmund, Germany) and PECAS (Calgary, Canada) are similar to UrbanSim in their approach in terms of their level of spatial disaggregation and the explicit modelling of individual agent behavior. Many previous studies have categorized and evaluated urban land-use/transport/environmental models based on attributes such as the level of disaggregation, the ability to incorporate dynamics, methodological approach and whether they are region specific or generic. Readers are especially directed to the reviews by Haase and Schwarz (2009), Agarwal et al. (2002) and Wegener (2004) for a more in-depth examination of the different model characteristics.

4.2 Spatial Modelling with Machine Learning

Conventional statistical models have stringent assumptions about the underlying data, which can be limiting if the assumptions do not hold. One important assumption in most classical statistical inference is based on the central limit theorem, which states that the probability distribution of the sum (or average) of independent and identically distributed

(IID) variables with finite variance approaches a normal distribution. Machine learning is a new computing-based statistical algorithm that requires minimal assumptions about the characteristics of the data or how they are associated with each other. Doherty et al. (2016, p. 30) describe machine learning as algorithms that 'enable machines to make decisions informed by data, where the machine has 'learned' to perform some task through exposure to training data'. The greater the volume of data supplied for training the model, the better prediction accuracy can be achieved from the test data.

There are two generic types of machine-learning techniques – supervised and unsupervised. The supervised version of machine learning includes data that are labeled, that is, each record has features and a label that is associated with some outcome. The objective of fitting the supervised model is to predict the outcome (label) of a test dataset that is not part of the data that have been used for training the model. In contrast, unsupervised machine learning uses data that are unlabeled. Therefore, the purpose of unsupervised learning is to discern patterns in data that are not known beforehand, such as in analyzing clusters of data. Often knowing what the clusters represent, or even the number of clusters, is valuable for building a supervised model for prediction. The following sections offer examples of the application of machine learning in spatial models.

Application of machine learning in energy modelling

Tracking energy consumption in the built environment to assess how our energy efficiency objectives are being met is an important aspect of planning (Holden and Norland 2005; Madlener and Sunak 2011). Several new models are now making use of machine learning to estimate energy consumption by training a large volume of data about the built environment and detailed information about climatic conditions. Robinson et al. (2017) estimated the commercial building energy footprint using built environment and activity characteristics such as square footage, number of floors, cooling degree days, heating degree days, number of employees, total hours open per week, months in use and food sales. The annual major fuel consumption was used as the predicted variable. Thirteen machine-learning algorithms were tested, including XGBoost, Bagging, MLP regressor, Random Forest regressor, KNN regressor, Ridge regressor and Lasso. The results show that gradient boosting regression models perform the best at predicting commercial building energy consumption. The trained model was validated based on an energy consumption dataset from New York City Local Law 84. Only five common building features explained a significant part of the energy consumption, which indicates that the data requirements for undertaking similar analyses for other cities might be lower.

Similar research on energy consumption in residential buildings was conducted by Tsanas and Xifara (2012). Eight building features, such as relative compactness, wall area, overall height and glazing area distribution were used as the predictor variables. A linear regression model and random forest model were tested to estimate heating load and cooling load. The results show that the machine-learning method is a feasible and accurate approach to estimate energy performance.

Machine learning for transportation

How to make a transportation system more efficient and accessible is a key mission for transportation planners (Kitamura 1990; Straatemeier 2008). An important aspect of transportation planning is understanding how people travel to different activity locations within the urban region. New technology now utilizes the use of GPS for tracking how different vehicles move in an urban region. For example, Zheng et al. (2008) used raw GPS data and supervised machine learning to determine travel mode choice. Their approach was designed around three analytical methods: the change point-based segmentation method; an inference model; and a post-processing component. Four inference models, including decision tree, Bayesian net, support vector machine and conditional random field, were tested with GPS data from 45 users over six months. The results indicated that this method can detect the transportation mode accurately and the decision-tree inference method offered the best outcome.

Another study, also incorporating GPS data, was conducted by Stenneth et al. (2011) to detect users' transportation modes. This study was based on data on the average accuracy of GPS coordinates, the average speed of vehicles, average heading change, average acceleration, distance to bus stops, distance to rail, and zip code-based indexing and pruning. The predicted modes included train, bus, car, walk, bike and stationary. Five machine-learning models were tested, including naive Bayes, Bayesian network, decision tree, random forest and multilayer perceptron. The results confirm that the machine-learning approach together with GPS and other data can offer high levels of accuracy in detecting transportation mode.

Machine learning for environmental planning

Air quality is an important environmental indicator that affects the quality of life and health of urban residents (Brunekreef and Holgate 2002). A significant aspect of air quality measurement is the concentration of suspended particulate matter in the air, particularly those that are less than 2.5 microns (PM2.5) and also those between 2.5 and 10 microns (PM10). Zickus et al. (2002) used machine-learning methods to predict PM10 concentrations in Helsinki, Finland. Four machine-learning algorithms were tested in this study, including logistic regression, decision tree, multivariate adaptive regression splines and neural network. The daily average of PM10 exceedances was used as the target variable. The predictor variables included wind speed, wind direction, pressure, relative humidity, temperature and cloudiness. The results show that three of the five models tested performed well in the test phase and could be used for predicting PM10 exceedances based on the predictor variables.

Another study by Li et al. (2016) proposed a novel spatiotemporal deep-learning based methodology to predict air quality. Three machine-learning models, including spatiotemporal artificial neural network (STANN), autoregression moving average (ARMA) and support vector regression (SVR) models, were compared with their own model. The deep-learning model their study proposed performed substantially better in predicting air quality when compared with the three machine-learning based models tested.

Measuring equity with machine learning models

Spatial modelling has been an important tool for identifying and understanding social equity issues (Korpi and Palme 1998). Jean et al. (2016) demonstrated a machine-learning based technique to estimate the variation in consumption expenditure and asset wealth in an urban region by analyzing high-resolution satellite imagery. A convolutional neural network model was used to identify image features. Labeled images with different categories were used as the training dataset. The model was trained to predict the nighttime light intensities with the help of data extracted from daytime satellite imagery. The mean cluster level values from a survey together with the image features from daytime satellite imagery were used to train a ridge regression model for estimating cluster-level expenditures or assets. The results show that a trained convolutional neural network can explain 75 percent of the variation in economic conditions across the region.

5. PLANNING URBAN FUTURES THROUGH SCENARIO DEVELOPMENT

Spatial models do not forecast future urban states; people do, based on their knowledge and intuition. Here we propose one approach that leverages multiple perspectives and expectations about the future from stakeholders to generate many scenarios. These scenarios can then be critically examined with the help of spatial models to determine what policies are necessary to reach the desired future outcomes.

Our approach begins with a framework for crowdsourcing rationally derived forecasts of aspects of the future from both experts and critical stakeholders. Consider an exercise with stakeholders and experts that is designed to anticipate the combined effects of five types of transitions. In this illustration, the group exercise coalesced around the following five significant changes that would impact the future urban state: (1) demographic shifts; (2) economic transformations; (3) changing lifestyles; (4) global climate change impacts; and (5) technological transitions (Table 8.1). The possible combinations of various expectations within each of the five categories of changes can lead to over 1500 scenarios. The advantage of generating a large volume of scenarios is that we are forced to consider possibilities that were previously outside our frame of reference. The problem with this approach is the difficulty in analyzing each of them, given the sheer number of scenarios.

Given that it would be impractical to deal with every scenario individually, a prudent approach would be to develop a framework for categorizing them into bins, thereby reducing the complexity. A simple and well-tested method for categorizing the scenarios is to place them on a two-dimensional matrix. In our case, we will use the two dimensions: (1) likelihood (probability of occurrence); and (2) impact (potential resources needed to manage/mitigate). All scenarios would include one possibility from each of the five different categories of transitions discussed above. Figure 8.1 explains how the scenarios with five forecasted events, one from each category, could then be displayed on the two-dimensional matrix. However, to get to this stage, we will need to assign an indicator of likelihood and impact for each scenario.

Since the forecasted events for each category were all crowdsourced from stakeholders and experts, the rationally defensible way to assign likelihood and impact would be to ask them for these values. So, each participant scores every possible forecasted event in all five categories (not just those they have proposed) with two values: (1) probability of occurrence expressed in percentage; and (2) an impact score on a scale of 1 to 5, 5 being the highest impact. For simplicity, we dispense with positive or negative impacts and only score on the overall resources needed to mitigate or manage the event.

Once all forecasted events have received their scores on probability and impact from all participants, standard statistical tools can be used to ascertain how these can be aggregated to reflect overall likelihood for a scenario (that combines one event from each

Table 8.1 Crowdsourcing potential future states by transition categories

Demographic shifts	Economic transformation	Changing lifestyle choices for housing	GCC impacts	Technological transformations
High ethnic diversity with around 50% Hispanic households	A large service and hospitality industry, declining manufacturing	High density, low-rise, clustered around retail/ commercial at the city fringe	Increasing cost of food, energy, and water	Improved solar technology, more prevalent use in new construction
Increasing median age of population	Increased economic concentration of nanotechnology and biotechnology industries	High-density, high-rise, clustered around historic downtown cores	Increasing heat and humidity from urban heat islands	Fossil fuel use in transportation falls to less than 50%
Single female-headed households increase to over 30%	Rapidly rising capital-to-labor ratio with few jobs for low-skilled workers	Low density, low-rise with leapfrog development	Massive in-migration from south of the border	High prevalence of virtual offices with improved networking and video streaming
Two-parent family households increase to over 80%	Military- and defense-related jobs concentrate in the region	High-density, mid-to-high rise around many new and old centers (polycentric)	Out-migration owing to extreme heat or coastal flooding	
Majority of the households are multi-ethnic/ multi-racial		A vertical city connected by individualized and flexible public transportation	Prolonged drought and water scarcity	

Note: 5 × 4 × 5 × 5 × 3 = 1500 minimum scenario possibilities (more if we consider combinations of two or more situations for one driver); These are then filtered to arrive at a small set of 10 to 15 scenarios that are plausible and important.

of categories). We illustrate this method using the same example, but with three transition categories, instead of five (Figure 8.2). It is easy to see that the procedure can be extended to five or more categories of transitions with the same set of models – but more categories of events will require increasingly large computing resources to process.

To ascertain the combined probability of any scenario with three separate events, each with their own mean probability (μ) and variance (σ2) (from data from all stakeholders), we use Bayesian statistics. The marginal probabilities of all events – conditional on other events in the scenario can be estimated using the following function:

$$P(S1) = P(\mu_{14}|\mu_{23} \cap \mu_{32}) = \frac{P(\mu_{14})\,P(\mu_{23}|\mu_{14})\,P(\mu_{32}|\mu_{14} \cap \mu_{23}}{P(\mu_{23})\,P(\mu_{32}|\mu_{23})} \tag{8.1}$$

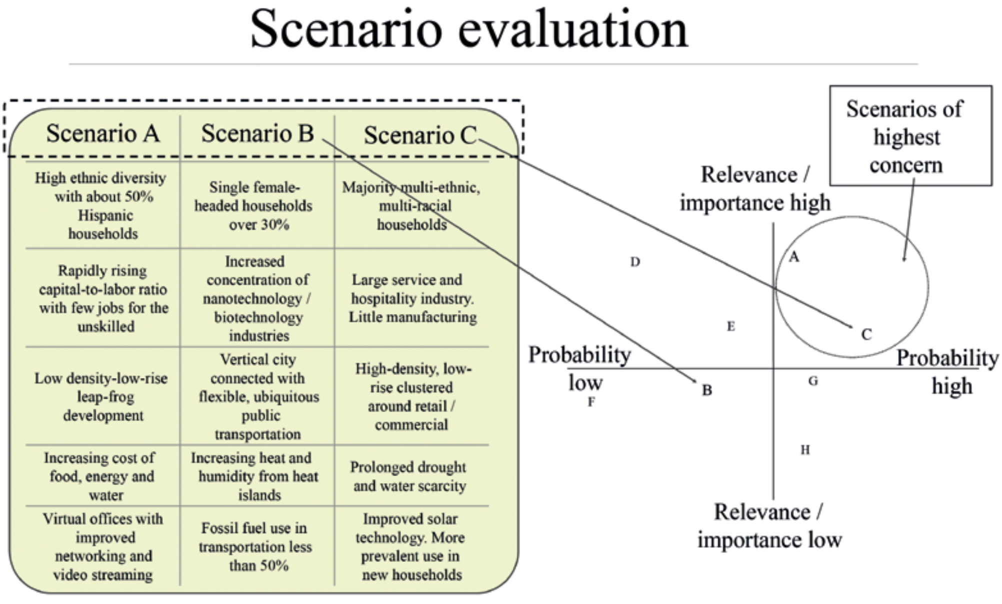

Figure 8.1 *Example of scenario analysis using a two-dimensional matrix*

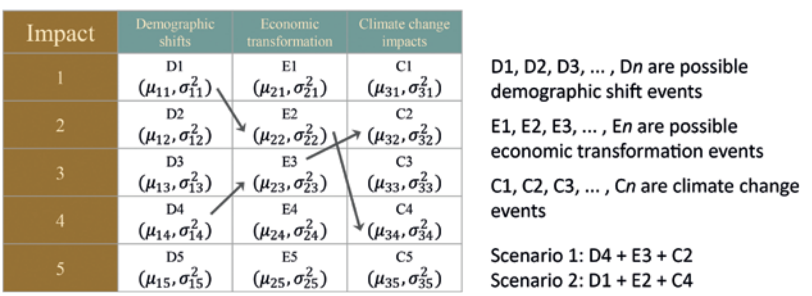

Figure 8.2 *Assigning likelihood and impact for each scenario based on three transition event categories*

where: $P(S1)$ is the probability of scenario 1, which is based on the probability of $D4$ (μ14) occurring conditional upon both $E3$ and $C2$ occurring (μ23 ∩ μ32).

This function will get larger when additional categories of events are included. Therefore, it is necessarily a computationally intensive task. However, several libraries are available in R and Python that can handle the complexity of deriving conditional probabilities for a series of events.

The estimation of joint impacts is more straightforward. We can use an additive or multiplicative function together with weights (if any). So, the impact for scenario 1 can be calculated as:

$$\text{Impact of } S1 = [w_d \times D_4 \times w_e \times E_3 \times w_c \times C_2 \tag{8.2}$$

where w_d, w_e, and w_c are weights that could be assigned to each category of event and D_4, E_3 C_2 are impact scores for events *D*4, *E*3, and *C*2, respectively. For example, if *D*4 is assigned a score of 4, *E*3 is assigned 3, and *C*2 is assigned 2, for their individual impacts respectively, then scenario 1 would have an impact score of 24 (with all weights assumed to be 1). The impact scores can be normalized by dividing by 125, which is the maximum possible impact with no weights. Normalization is not necessary in this context if we can derive a rational way to segment out higher and lower impact scores. The objective is to place these scores along one of the two dimensions shown in Figure 8.1.

When all the scenarios are visualized along the two-dimensional matrix shown in Figure 8.1, they can help us focus on certain groups of scenarios. Figure 8.3 shows different planning approaches that would be appropriate for each of the four quadrants in our two-dimensional matrix. The high impact but rare scenarios are difficult to address using local resources. These are either ignored completely or left to the purview of national and international entities that can quickly deploy emergency management resources. The low-impact and low-probability scenarios can be safely ignored. The high-probability but low-impact scenarios need to be evaluated to see if these have longer-term effects that need to be managed. The group of scenarios that require the most attention is the high-impact and high-probability scenarios. These could be further disaggregated by the type of impact (for example, slowly evolving versus sudden pulses) to develop appropriate planning strategies for managing and mitigating their severity. The scenarios do not provide information about the spatial variation in impacts that could be felt. Spatial models are critical for assessing the spatial dimensions of the impacts, which would be a valuable input for planning.

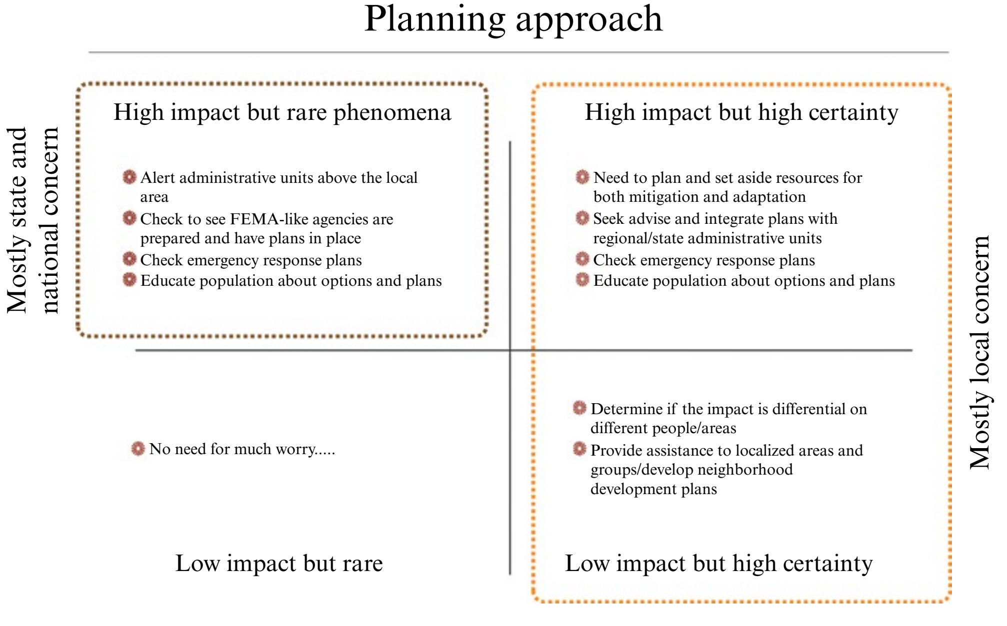

Figure 8.3 Grouping scenarios for planning approaches

6. SPATIAL MODELS AND FORECASTING IN THE NEW ERA

The confluence of several technological and social trends is revolutionizing the way we interact with our physical surroundings and with each other. High-speed communication networks have transformed many of the city functions such as transport, service delivery and emergency management. Devices that sense locations, see objects and measure the ambient environment (temperature, noise, air quality and other activities) are everywhere. Cell phone records can track the flow of people and crowds through different parts of the city. Data from social media such as Twitter and Instagram can provide rich information about the most significant events that are happening in real time. The impact on city form and function from a combination of different transformative technologies that are just beginning to be felt is difficult to ascertain. Carefully designed spatial models that take into account the actual behavior of multiple agents can provide a more accurate account of what lies ahead for urban residents.

Urban spatial models are designed to portray the future state of a region by showing how its essential functions are going to change and how these functions may be accommodated in space. Where people live and work, how they travel, what buildings they use, and how they spend their leisure are important questions for determining the sustainability of that region. By projecting social behavior and conditions of urban life into the future, urban models also allow us to estimate the amount of energy, water and open space that will be needed. In addition, they can inform us about the quality of air, water and land we will inherit in that future. However, urban models are not omniscient; they only project what we can imagine. It is only our imagination and our ability to form mental models that ultimately allow us to contemplate and envision urban sustainability. Urban models are the tools to make our mental models correspond with reality.

Regardless of the new tools, advanced technologies and new sources of spatial data, the outcome of spatial models is still subject to particularities of the people and places. Also, analytical results need to be legitimized through extended peer communities. Expanding sources and volumes of data creates new risks and concerns for privacy that need to be balanced with efficiency and usability. High connectivity of urban systems offers systemwide benefits in improved efficiency, but also systemwide vulnerabilities if small failures in one part of the system lead to a cascading effect infecting other parts. Therefore, it would be unwise to believe that automated systems would dispense with human interventions and oversight. Only thoughtful management systems with human-in-the-loop oversight mechanisms, together with increased engagement from stakeholders, can both offer the benefits of new technologies and avoid their pitfalls.

7. CONCLUSIONS

This chapter covered an extensive analytical territory to provide an overview of the methods, applications and the changing context for spatial analysis together with some strategies for forecasting future states. It offered a framework for developing future forecasts using a participatory approach, which can be used to guide the spatial modelling of urban futures. The principal objective was to situate current modelling efforts within the innovations in data acquisition, preparation, storage and the new analytical tools that are spurring investments in smart-city technologies. The chapter also contextualized the political and socioeconomic ethos of current urban planning, which is characterized as the post-normal period, where agreement on values, problems and the strategies to address them is rare. Planning and social modelling in this era is more grounded and the solutions are specific to the time, place and the constituencies served. The solutions are arrived at through an iterative process of converging interests that results in a particular agreement for addressing the problem. These agreements are then legitimized by the extended peer communities of experts, community leaders and stakeholders.

While the problems are tackled through a process that generates particularistic solutions specific to the context, they are often universal. The specter of climate change impacts, rising inequality, inadequate provision of basic human necessities such as housing, healthcare, food, clean air and clean water, have raised concerns in far-flung communities across the globe. There is a critical need for expanding the means through which we understand the individualized and collective human conditions in different places and situations. Our new tools for extracting large amounts of data about human behavior and motivations and the environment, as well as for understanding these patterns within the data, are an important part of formulating strategies. Spatial modelling and forecasting helps to focus our attention on the spatially heterogeneous characteristics of people and places as well as their trends. They are often necessary, but by no means sufficient, for designing plans for a better urban future. Ultimately, all affected constituencies have to be engaged to create such a future state that meets everyone's aspirations.

REFERENCES

Agarwal, C., Green, G.M., Grove, J.M., Evans, T.P. and Schweik, C.M. (2002), *A Review and Assessment of Land-Use Change Models: Dynamics of Space, Time, and Human Choice*, Newton Square, PA: US Department of Agriculture, Forest Service, Northeastern Research Station.

Anselin, L. (1995), 'Local indicators of spatial association – LISA', *Geographical Analysis*, **27** (2), 93–115.

Anselin, L. and Gallo, J.L. (2006), 'Interpolation of air quality measures in hedonic house price models: spatial aspects', *Spatial Economic Analysis*, **1** (1), 31–52.

Anselin, L. and Rey, S.J. (2014), *Modern Spatial Econometrics in Practice: A Guide to GeoDa, GeoDaSpace and PySAL*, Chicago, IL: GeoDa Press.

Asgary, A., Klosterman, R. and Razani, A. (2007), 'Sustainable urban growth management using What-if?', *International Journal of Environmental Resources*, **1** (3), 218–30.

Axinn, W.G. and Ghimire, D.J. (2011), 'Social organization, population, and land use', *American Journal of Sociology*, **117** (1), 209–58.

Bailly, S., Meyfroidt, G. and Timsit, J.-F. (2018), 'What's new in ICU in 2050: big data and machine learning', *Intensive Care Medicine*, **44** (9), 1524–7.

Bifulco, G., Galante, F., Pariota, L., Spena, M.R. and Del Gais, P. (2014), 'Data collection for traffic and drivers' behaviour studies: a large-scale survey', *Procedia-Social and Behavioral Sciences*, **111** (February), 721–30.

Blanchet, J. and Lehning, M. (2010), 'Mapping snow depth return levels: smooth spatial modelling versus station interpolation', *Hydrology and Earth System Sciences*, **14** (12), 2527–44.

Brunekreef, B. and Holgate, S.T. (2002), 'Air pollution and health', *The Lancet*, **360** (9341), 1233–42.

Cao, J., Huang, Z. and Yang, Y. (2015), 'Spatial-aware multimodal location estimation for social images', in *Proceedings of the 23rd ACM International Conference on Multimedia*, New York: ACM, pp. 119–28.

Chaudhuri, G. and Clarke, K. (2013), 'The SLEUTH land use change model: a review', *Environmental Resources Research*, **1** (1), 88–105.

Cheng, Q. (1999), 'Spatial and scaling modelling for geochemical anomaly separation', *Journal of Geochemical Exploration*, **65** (3), 175–94.

Collins, M.G., Steiner, F.R. and Rushman, M.J. (2001), 'Land-use suitability analysis in the United States: historical development and promising technological achievements', *Environmental Management*, **28** (5), 611–21.

Cukier, K. (2010), 'Data, data everywhere: a special report on managing information', *The Economist*, 27 February.

Curtis, A., Blackburn, J.K., Widmer, J.M. and Morris Jr, J.G. (2013), 'A ubiquitous method for street scale spatial data collection and analysis in challenging urban environments: mapping health risks using spatial video in Haiti', *International Journal of Health Geographics*, **12** (1), 21.

Dalton, L.C. (1986), 'Why the rational paradigm persists – the resistance of professional education and practice to alternative forms of planning', *Journal of Planning Education and Research*, **5** (3), 147–53.

De Mauro, A., Greco, M. and Grimaldi, M. (2015), 'What is big data? A consensual definition and a review of key research topics', in *AIP Conference Proceedings*, **1644**, AIP, Madrid, 5–8 September, pp. 97–104.

Deal, B. and Schunk, D. (2004), 'Spatial dynamic modelling and urban land use transformation: a simulation approach to assessing the costs of urban sprawl', *Ecological Economics*, **51** (1), 79–95.

Doherty, C., Camiña, S., White, K. and Orenstein, G. (2016), *The Path to Predictive Analytics and Machine Learning*, Sebastopol, CA: O'Reilly Media.

Engelen, G., White, R. and Uljee, I. (1997), 'Integrating constrained cellular automata models, GIS and decision support tools for urban planning and policy-making', in H.P.J. Timmermans (ed.), *Decision Support Systems in Urban Planning*, London: E & FN Spon, pp. 125–55.

Fang, S., Gertner, G.Z., Sun, Z. and Anderson, A.A. (2005), 'The impact of interactions in spatial simulation of the dynamics of urban sprawl', *Landscape and Urban Planning*, **73** (4), 294–306.

Friedmann, J. (1978), 'The epistemology of social practice', *Theory and Society*, **6** (1), 75–92.

Funtowicz, S. and Ravetz, J. (2003), 'Post-normal science', in International Society for Ecological Economics (ed.), *Internet Encyclopedia of Ecological Economics*, accessed 19 September 2019 at http://isecoeco.org/pdf/pstnormsc.pdf.

Gandomi, A. and Haider, M. (2015), 'Beyond the hype: big data concepts, methods, and analytics', *International Journal of Information Management*, **35** (2), 137–44.

Getis, A. and Ord, J.K. (1992), 'The analysis of spatial association by use of distance statistics', *Geographical Analysis*, **24** (3), 189–206.

Glaeser, E.L., Kominers, S.D., Luca, M. and Naik, N. (2018), 'Big data and big cities: the promises and limitations of improved measures of urban life', *Economic Inquiry*, **56** (1), 114–37.

Godet, M. and Roubelat, F. (1996), 'Creating the future: the use and misuse of scenarios', *Long Range Planning*, **29** (2), 164–71.

Goldner, W. (1971), 'The Lowry model heritage', *Journal of the American Institute of Planners*, **37** (2), 100–110.

Guhathakurta, S., Zhang, G., Chen, G., Burnette, C. and Sepkowitz, I. (2019), 'Mining social media to measure neighborhood quality in the city of Atlanta', *International Journal of E-Planning Research (IJEPR)*, **8** (1), 1–18.

Haase, D. and Schwarz, N. (2009), 'Simulation models on human-nature interactions in urban landscapes: a review including spatial economics, system dynamics, cellular automata and agent-based approaches', *Living Reviews in Landscape Research*, **3** (2), 1–45.

Hauff, C. and Houben, G.-J. (2012), 'Geo-location estimation of flickr images: social web based enrichment', in R. Baeza-Yates, A.P. de Vries, H. Zaragoza, B. Barla Cambazoglu, V. Murdock, R. Lempel, et al. (eds), *Advances in Information Retrieval: 34th European Conference on IR Research, ECIR 2012, Barcelona, Spain, April 1–5, 2012, Proceedings*, Berlin and Heidelberg: Springer, pp. 85–96.

Holdaway, M.R. (1996), 'Spatial modelling and interpolation of monthly temperature using kriging', *Climate Research*, **6** (3), 215–25.

Holden, E. and Norland, I.T. (2005), 'Three challenges for the compact city as a sustainable urban form: household consumption of energy and transport in eight residential areas in the Greater Oslo Region', *Urban Studies*, **42** (12), 2145–66.

Houghton, J.T., Ding, Y., Griggs, D.J., Noguer, M., Van Der Linden, P.J., Dai, X., et al. (eds) (2001), *IPCC, 2001: Climate Change 2001: The Scientific Basis. Contribution of Working Group 1 to the Third Assessment Report of the International Panel on Climate Change*, Cambridge: Cambridge University Press.

Hu, S., Cheng, Q., Wang, L. and Xu, D. (2013), 'Modelling land price distribution using multifractal IDW interpolation and fractal filtering method', *Landscape and Urban Planning*, **110** (February), 25–35.

Hu, Z. and Lo, C.P. (2007), 'Modelling urban growth in Atlanta using logistic regression', *Computers, Environment and Urban Systems*, **31** (6), 667–88.

Huang, B., Wu, B. and Barry, M. (2010), 'Geographically and temporally weighted regression for modelling spatio-temporal variation in house prices', *International Journal of Geographical Information Science*, **24** (3), 383–401.

Huang, J., Lu, X.X. and Sellers, J.M. (2007), 'A global comparative analysis of urban form: applying spatial metrics and remote sensing', *Landscape and Urban Planning*, **82** (4), 184–97.

Hunt, J.D. and Abraham, J.E. (2005), 'Design and implementation of PECAS: a generalised system for allocating economic production, exchange and consumption quantities', in M.E.H. Lee-Gosselin and S.T. Doherty (eds), *Integrated Land-Use and Transportation Models: Behavioural Foundations*, Bingley: Emerald, pp. 253–73.

Jardak, C., Mähönen, P. and Riihijärvi, J. (2014), 'Spatial big data and wireless networks: experiences, applications, and research challenges', *IEEE Network*, **28** (4), 26–31.
Jean, N., Burke, M., Xie, M., Davis, W.M., Lobell, D.B. and Ermon, S. (2016), 'Combining satellite imagery and machine learning to predict poverty', *Science*, **353** (6301), 790–94.
Karnieli, A., Gilad, U., Ponzet, M., Svoray, T., Mirzadinov, R. and Fedorina, O. (2008), 'Assessing land-cover change and degradation in the Central Asian deserts using satellite image processing and geostatistical methods', *Journal of Arid Environments*, **72** (11), 2093–105.
Kitamura, R. (1990), 'Panel analysis in transportation planning: an overview', *Transportation Research Part A: General*, **24** (6), 401–15.
Kitchin, R. (2014), 'The real-time city? Big data and smart urbanism', *GeoJournal*, **79** (1), 1–14.
Klosterman, R.E. (1999), 'The What If? collaborative planning support system', *Environment and Planning B: Planning and Design*, **26** (3), 393–408.
Klosterman, R.E. (2008), 'A new tool for a new planning: The What If? planning support system', in R.K. Brail (ed.), *Planning Support Systems for Cities and Regions*, Cambridge, MA: Lincoln Institute of Land Policy, pp. 85–100.
Korpi, W. and Palme, J. (1998), 'The paradox of redistribution and strategies of equality: welfare state institutions, inequality, and poverty in the western countries', *American Sociological Review*, **63** (5), 661–87.
Krieger, N., Williams, D.R. and Moss, N.E. (1997), 'Measuring social class in US public health research: concepts, methodologies, and guidelines', *Annual Review of Public Health*, **18** (1), 341–78.
Landis, J.D. (1994), 'The California urban futures model: a new generation of metropolitan simulation models', *Environment and Planning B: Planning and Design*, **21** (4), 399–420.
Landis, J.D. (1995), 'Imagining land use futures: applying the California urban futures model', *Journal of the American Planning Association*, **61** (4), 438–457.
Lee, J.-G. and Kang, M. (2015), 'Geospatial big data: challenges and opportunities', *Big Data Research*, **2** (2), 74–81.
Lee, K., Ganti, R.K., Srivatsa, M. and Liu, L. (2014), 'When twitter meets foursquare: tweet location prediction using foursquare', in *Proceedings of the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, London: Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering (ICST), pp. 198–207.
Lee, S.-J., Wentz, E.A. and Gober, P. (2010), 'Space-time forecasting using soft geostatistics: a case study in forecasting municipal water demand for Phoenix, Arizona', *Stochastic Environmental Research and Risk Assessment*, **24** (2), 283–95.
Leeuw, E. de and Collins, M. (1997), 'Data collection methods and survey quality: an overview', in L. Lyberg, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz and D Trewin (eds), *Survey Measurement and Process Quality*, New York: Jouhn Wiley & Sons, pp. 197–220.
Li, X., Peng, L., Hu, Y.,Shao, J. and Chi, T. (2016), 'Deep learning architecture for air quality predictions', *Environmental Science and Pollution Research*, **23** (22), 22408–17.
Lowry, I.S. (1964), *A Model of Metropolis*, Santa Monica, CA: Rand Corporation.
Lu, G.Y. and Wong, D.W. (2008), 'An adaptive inverse-distance weighting spatial interpolation technique', *Computers & Geosciences*, **34** (9), 1044–55.
Maack, J.N. (2001), 'Scenario analysis: a tool for task managers', in R.A. Krueger, M. Casey, J. Donner, S. Kirsch and J.N. Maack (eds), *Social Analysis Selected Tools and Techniques*, Washington, DC: World Bank, pp. 62–87.
Madlener, R. and Sunak, Y. (2011), 'Impacts of urbanization on urban structures and energy demand: what can we learn for urban energy planning and urbanization management?', *Sustainable Cities and Society*, **1** (1), 45–53.
Malczewski, J. (2004), 'GIS-based land-use suitability analysis: a critical overview', *Progress in Planning*, **62** (1), 3–65.
McCarty, J.J., Canziani, O.F., Leary, N.A., Dokken, D.J. and White, K.S. (2001), *Climate Change 2001: Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Third Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge: Cambridge University Press.
McHarg, I.L. and Mumford, L. (1969), *Design with Nature*, New York: American Museum of Natural History.
McMillen, D.P. (2004), 'Geographically weighted regression: the analysis of spatially varying relationships', *American Journal of Agricultural Economics*, **86** (2), 554–6.
Miller, W., Collins, M.G., Steiner, F.R. and Cook, E. (1998), 'An approach for greenway suitability analysis', *Landscape and Urban Planning*, **42** (2), 91–105.
Mitas, L. and Mitasova, H. (1999), 'Spatial interpolation', in P. Longley, M. Goodchild, D. Maguire and D. Rhind (eds), *Geographical Information Systems: Principles, Techniques, Management and Applications, Volume 1 Principles and Technical Issues*, London: Wiley, pp. 481–9).
Mohammadi, E., Thelwall, M., Kwasny, M. and Holmes, K.L. (2018), 'Academic information on Twitter: a user survey', *PloS One*, **13** (5), e0197265.

Mosadeghi, R., Warnken, J., Tomlinson, R. and Mirfenderesk, H. (2015), 'Comparison of fuzzy-AHP and AHP in a spatial multi-criteria decision making model for urban land-use planning', *Computers, Environment and Urban Systems*, **49** (January), 54–65.
Naik, N., Philipoom, J., Raskar, R. and Hidalgo, C. (2014), 'Streetscore-predicting the perceived safety of one million streetscapes', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Boston, MA: Institute of Electrical and Electronics Engineers, pp. 779–85.
Obermeyer, Z. and Emanuel, E.J. (2016), 'Predicting the future – big data, machine learning, and clinical medicine', *New England Journal of Medicine*, **375** (13), 1216–19.
Oliveira, S., Oehler, F., San-Miguel-Ayanz, J., Camia, A. and Pereira, J.M.C. (2012), 'Modelling spatial patterns of fire occurrence in Mediterranean Europe using multiple regression and random forest', *Forest Ecology and Management*, **275** (July), 117–29.
Ord, J.K. and Getis, A. (1995), 'Local spatial autocorrelation statistics: distributional issues and an application', *Geographical Analysis*, **27** (4), 286–306.
Putman, S. and Chan, S.-L. (2001), 'The METROPILUS planning support system: urban models and GIS', in R. Brail and R. Klosterman (eds), *Planning Support Systems*, Redlands, CA: Esri Press, pp. 99–128.
Putman, S.H. (1983), *Integrated Urban Models: Policy Analysis of Transportation and Land Use*, London: Pion.
Rittel, H.W. and Webber, M.M. (1973), 'Dilemmas in a general theory of planning', *Policy Sciences*, **4** (2), 155–69.
Robinson, C., Dilkina, B., Hubbs, J., Zhang, W., Guhathakurta, S., Brown, M.A. et al. (2017), 'Machine learning approaches for estimating commercial building energy consumption', *Applied Energy*, **208** (December), 889–904.
Romesburg, C. (2004), *Cluster Analysis for Researchers*, Lulu.com.
Rovatti, R., Guerrieri, R. and Borgatti, M. (1998), 'A geometric approach to maximum-speed n-dimensional continuous linear interpolation in rectangular grids', *IEEE Transactions on Computers*, **47** (8), 894–9.
Sambridge, M., Braun, J. and McQueen, H. (1995), 'Geophysical parametrization and interpolation of irregular data using natural neighbours', *Geophysical Journal International*, **122** (3), 837–57.
Santelmann, M., Freemark, K., White, D., Nassauer, J., Clark, M., Danielson, B., et al. (2001), 'Applying ecological principles to land-use decision making in agricultural watersheds', in V.H. Dale and R.A. Haeuber (eds), *Applying Ecological Principles to Land Management*, New York: Springer, pp. 226–52.
Seo, C., Thorne, J.H., Hannah, L. and Thuiller, W. (2008), 'Scale effects in species distribution models: implications for conservation planning under climate change', *Biology Letters*, **5** (1), 39–43.
Shekhar, S. (2012), 'Spatial big data challenges', paper presented at the Keynote at ARO/NSF Workshop on Big Data at Large: Applications and Algorithms, Durham, NC, 14 June.
Silva, E.A. and Clarke, K.C. (2002), 'Calibration of the SLEUTH urban growth model for Lisbon and Porto, Portugal', *Computers, Environment and Urban Systems*, **26** (6), 525–52.
Stein, M.L. (2012), *Interpolation of Spatial Data: Some Theory for Kriging*, Berlin: Springer Science & Business Media.
Steinitz, C., Arias, H., Bassett, S., Flaxman, M., Goode, T., Maddock III, T., et al. (2003), *Alternative Futures for Changing Landscapes: The Upper San Pedro River Basin in Arizona and Sonora*, Washington, DC: Island Press.
Stenneth, L., Wolfson, O., Yu, P.S. and Xu, B. (2011), 'Transportation mode detection using mobile phones and GIS information', in *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, New York: ACM, pp. 54–63.
Straatemeier, T. (2008), 'How to plan for regional accessibility?', *Transport Policy*, **15** (2), 127–37.
Tsai, C.-H. and Chen, C.-W. (2011), 'The establishment of a rapid natural disaster risk assessment model for the tourism industry', *Tourism Management*, **32** (1), 158–71.
Tsanas, A. and Xifara, A. (2012), 'Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools', *Energy and Buildings*, **49** (June), 560–67.
Vatsavai, R.R., Ganguly, A., Chandola, V., Stefanidis, A., Klasky, S. and Shekhar, S. (2012), 'Spatiotemporal data mining in the era of big spatial data: algorithms and applications', in *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data*, New York: ACM, pp. 1–10.
Verburg, P.H. and Veldkamp, A. (2005), 'Introduction to the special issue on spatial modelling to explore land use dynamics', *International Journal of Geographical Information Science*, **19** (2), 99–102.
Vittek, M., Brink, A., Donnay, F., Simonetti, D. and Desclée, B. (2014), 'Land cover change monitoring using Landsat MSS/TM satellite image data over West Africa between 1975 and 1990', *Remote Sensing*, **6** (1), 658–76.
Waddell, P. (2002), 'UrbanSim: modelling urban development for land use, transportation, and environmental planning', *Journal of the American Planning Association*, **68** (3), 297–314.
Wegener, M. (2004), 'Overview of land-use transport models', in D.A. Hensher, K.J. Button, K.E. Haynes and P.R. Stopher (eds), *Handbook of Transport Geography and Spatial Systems*, vol. 5, Bingley: Emerald, pp. 127–46.

Wu, C. and Murray, A.T. (2005), 'A cokriging method for estimating population density in urban areas', *Computers, Environment and Urban Systems*, **29** (5), 558–79.

Yu, B., Shu, S., Liu, H., Song, W., Wu, J., Wang, L. et al. (2014), 'Object-based spatial cluster analysis of urban landscape pattern using nighttime light satellite images: a case study of China', *International Journal of Geographical Information Science*, **28** (11), 2328–55.

Zheng, Y., Liu, L., Wang, L. and Xie, X. (2008), 'Learning transportation mode from raw GPS data for geographic applications on the web', in *Proceedings of the 17th International Conference on World Wide Web*, New York: ACM, pp. 247–56.

Zickus, M., Greig, A.J. and Niranjan, M. (2002), 'Comparison of four machine learning methods for predicting PM10 concentrations in Helsinki, Finland', *Water, Air and Soil Pollution: Focus*, **2** (5), 717–29.