Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Preventive Veterinary Medicine 122 (2015) 213–220

Contents lists available at ScienceDirect

Preventive Veterinary Medicine

journal homepage: www.elsevier.com/locate/prevetmed

Spatial and temporal epidemiological analysis in the Big Data era

Dirk U. Pfeiffer*, Kim B. Stevens

Veterinary Epidemiology, Economics & Public Health Group, Department of Production & Population Health, Royal Veterinary College, London, UK

ARTICLE INFO

Article history:
Received 3 February 2015
Received in revised form 27 May 2015
Accepted 31 May 2015

Keywords:
Data science
Exploratory analysis
Internet of Things
Modelling
Multi-criteria decision analysis
Spatial analysis
Visualisation

ABSTRACT

Concurrent with global economic development in the last 50 years, the opportunities for the spread of existing diseases and emergence of new infectious pathogens, have increased substantially. The activities associated with the enormously intensified global connectivity have resulted in large amounts of data being generated, which in turn provides opportunities for generating knowledge that will allow more effective management of animal and human health risks. This so-called *Big Data* has, more recently, been accompanied by the *Internet of Things* which highlights the increasing presence of a wide range of sensors, interconnected via the Internet. Analysis of this data needs to exploit its complexity, accommodate variation in data quality and should take advantage of its spatial and temporal dimensions, where available. Apart from the development of hardware technologies and networking/communication infrastructure, it is necessary to develop appropriate data management tools that make this data accessible for analysis. This includes relational databases, geographical information systems and most recently, cloud-based data storage such as Hadoop distributed file systems. While the development in analytical methodologies has not quite caught up with the *data deluge*, important advances have been made in a number of areas, including spatial and temporal data analysis where the spectrum of analytical methods ranges from visualisation and exploratory analysis, to modelling. While there used to be a primary focus on statistical science in terms of methodological development for data analysis, the newly emerged discipline of *data science* is a reflection of the challenges presented by the need to integrate diverse data sources and exploit them using novel data- and knowledge-driven modelling methods while simultaneously recognising the value of quantitative as well as qualitative analytical approaches. Machine learning regression methods, which are more robust and can handle large datasets faster than classical regression approaches, are now also used to analyse spatial and spatio-temporal data. Multi-criteria decision analysis methods have gained greater acceptance, due in part, to the need to increasingly combine data from diverse sources including published scientific information and expert opinion in an attempt to fill important knowledge gaps. The opportunities for more effective prevention, detection and control of animal health threats arising from these developments are immense, but not without risks given the different types, and much higher frequency, of biases associated with these data.

1. Introduction

Economic and technological developments in the last 50 years have led to global eco-social system changes that greatly facilitate the emergence and spread of infectious diseases in both animals and humans. This represents a major challenge for the management of infectious disease risks and is likely to require a paradigm shift in analytical approaches rather than an evolution of existing ones. This change in approach is reflected in the widespread recognition of the need to adopt inter- and transdisciplinary approaches in risk research and management. In addition, the digital revolution has provided major opportunities with respect to data collection and analysis. This has now evolved into the Internet of Things where everyday objects are connected through information networks, allowing them to send and receive data (Anon., 2014b; Kamel Boulos and Al-Shorbaji, 2014). Related to this, is the so-called *Industry 4.0* (a collective term for technologies and concepts of value chain organisation; (Lee et al., 2014)), which reflects a vision for how the industrial sector may respond to the tight integration between the physical and digital world through the implementation of smart value chains.

* Corresponding author at: Veterinary Epidemiology, Economics & Public Health Group, Dept. of Production & Population Health, Royal Veterinary College, Hawkshead Lane, Hatfield, Hertfordshire, AL97TA, United Kingdom. Fax: +44 1707 666574.
E-mail address: pfeiffer@rvc.ac.uk (D.U. Pfeiffer).

http://dx.doi.org/10.1016/j.prevetmed.2015.05.012
0167-5877/

The concepts of smart health (Solanas et al., 2014), mHealth (Istepanian et al., 2004) and eHealth (Eysenbach, 2001) can be seen as the starting point for these developments and, together with the recent increase in popularity, and availability, of wearable sensors, have boosted the development of associated technologies. However, these sensors, other measurement devices and data sources are of limited use if the raw data they generate are not converted into information that can inform decision making, which has led to the need for suitable data management and analytical methods that can handle the resulting large, heterogeneous datasets.

In animal health in general, and veterinary epidemiology specifically, the established methodological frameworks provide guidance for research of cause-effect relationships based on data generated through *a priori* designed field and laboratory studies. This review explores recent developments, and future directions, for spatial and temporal analysis in support of managing complex animal health problems. We begin this review from a broader perspective by focussing on the developments that have led to the data revolution and its impact on the health sciences. We then discuss how the new scientific discipline of data science has been established to tackle the analytical challenges and opportunities resulting from the data revolution. From this wider analytical context, we then focus on the specific developments in spatio-temporal epidemiological data analysis resulting from the data revolution.

2. Data revolution: from the Internet via Big Data to the Internet of Things

Scientific approaches aimed at improving our understanding of the complexity of the systems of which animal and human diseases form a part, usually involve data collection. However, the way in which data are generated has changed radically over the last 30 years, mainly as a result of the emergence of electronic methods for measuring, recording, storing and distributing data. As part of this development, the Internet now forms the backbone of a globally-reaching information network. The drivers behind the data revolution have been multiple, and early on were dominated by defence, public safety and scientific interests. Only once commercial companies such as Google (https://www.google.com), Amazon (http://www.amazon.com) and Facebook (https://www.facebook.com) were able to demonstrate, during the last 10 years, the potential for commercial exploitation, did the data revolution truly take off. There are also now increasing concerns in relation to potential abuse of Big Data (Schadt, 2012; Anon., 2014a).

Mayer-Schönberger and Cukier (2014) define Big Data as '*The ability of society to harness information in novel ways to produce useful insights or goods and services of significant value*' and '*. . .things one can do at a large scale that cannot be done at a smaller one, to extract new insights or create new forms of value*.' Big data are generally characterised by 3Vs: volume (relative magnitude of dataset), velocity (rate at which new data are generated) and variety (heterogeneous structure of dataset [e.g. text, video, audio]) (Gandomi and Haider, 2015). A fourth 'v' frequently used to describe Big Data is veracity which acknowledges the inherent uncertainty frequently associated with, in particular, web-based Big Data and the corresponding need for analytical approaches that are able to account for this unreliability (Gandomi and Haider, 2015). In addition, the business community has added a fifth 'v'; *value*.

Traditional database management systems based on tabular or relational data management structures are not suited to dealing with Big Data as most of it is unstructured. Cloud-based data storage using the Apache Hadoop® distributed file system (http://hadoop.apache.org; last accessed April 2015) has been developed to allow efficient management of such data (O'Driscoll et al., 2013; Fernández et al., 2014). A data mining approach was used to explore the use of search term data for prediction of flu trends (Ginsberg et al., 2009) based on the assumption that changes in information and communication patterns on the Internet can act as early warning of changes in population health (Wilson and Brownstein, 2009). This resulted in the development of the search-term surveillance system, Google Flu Trends (GFT; http://www.google.org/flutrend, last accessed April 2015). By combining data-mining of Google search queries and statistical modelling, GFT provides a baseline indicator of the trend or changes in the rate of influenza, thereby providing estimates of weekly regional US influenza activity with a reporting lag of only one day compared with the 1–2 week delays associated with the Centers for Disease Control and Prevention (CDC) Influenza Sentinel Provider Surveillance reports (Ginsberg et al., 2009). However, the results generated by this algorithm have been the subject of controversy as predictions were incorrect at specific time points when they particularly mattered (Butler, 2013; Lazer et al., 2014). The fact remains though, that the relative immediacy of web-based surveillance systems allows for much quicker targeting of infection hot-spots in pandemic situations, as was done by companies such as Google, in the recent influenza H1N1 crisis (Chew and Eysenbach, 2010; Signorini et al., 2011; St Louis and Zorlu, 2012).

Although search-term surveillance systems such as GFT are currently best suited to track disease activity in developed countries – the system requires large populations of web-search users in order to be most effective (Carneiro and Mylonakis, 2009) and a robust existing surveillance system to provide data for calibration (Wilson et al., 2009), – retrospective analysis of Google Trend's search frequency for the term 'Ebola', in the developing countries of Guinea, Liberia and Sierra Leone, showed a moderate-to-high correlation with epidemic curves for the outbreak in those countries (Milinovich et al., 2015) suggesting that web-based surveillance systems have the potential to be used as early-warning systems in developing, as well as in developed, countries.

However, systems which mine secondary (e.g. news reports) rather than primary web-based data sources (e.g. search queries) are possibly better suited for disease surveillance in developing countries. Examples of such systems include BioCaster (http://biocaster.nii.ac.jp, last accessed April 2015; Collier et al., 2008), EpiSPIDER (Tolentino et al., 2007; Keller et al., 2009), HealthMap (http://www.healthmap.org, last accessed April 2015; Brownstein et al., 2008; Freifeld et al., 2008; Brownstein et al., 2009; Keller et al., 2009; Brownstein et al., 2010), ProMED-mail (http://www.promedmail.org, last accessed April 2015; Cowen et al., 2006; Tolentino et al., 2007; Zeldenrust et al., 2008) and Canada's Global Public Health Intelligence Network (GPHIN) (Mykhalovskiy and Weir, 2006). The value of such systems for flagging potential health threats is highlighted by the fact that GPHIN identified the 2002 severe acute respiratory syndrome (SARS) outbreak in Guangdong Province, China, more than two months before the World Health Organisation's (WHO) official announcement (Mykhalovskiy and Weir, 2006). Similarly, HealthMap identified news stories reporting a strange fever in Guinea nine days before official notification of the 2014 West Africa Ebola outbreak (Milinovich et al., 2015).

Although the inadequate initial response by the international community to the 2014 Ebola outbreak has been highlighted by some as a failure of Big Data analytical approaches for purposes of early warning (Leetaru, 2014; Milinovich et al., 2015), the fact remains that the primary value of such systems currently lies in their ability to flag events that may warrant further investigation rather than acting as the primary surveillance system (Wilson and Brownstein, 2009; Hartley et al., 2013). As such, although web-based surveillance systems are still a long way from replacing traditional surveillance methods, they provide a useful complement to conventional approaches (Milinovich et al., 2014), to the extent that they have become an important component of the

influenza surveillance scene. For example, WHO's Global Outbreak Alert and Response Network (GOARN; http://www.who.int/csr/outbreaknetwork/en/, last accessed May 2015) use such data as part of their day-to-day surveillance activities (Grein et al., 2000; Heymann and Rodier, 2001) and are authorised to act on this information (Wilson et al., 2008). Moving from surveillance to delivery of health care, precision medicine aims to utilise Big Data for the purpose of optimising the use of diagnostic tools, therapeutics and preventive management (Anon., 2011; Collins and Varmus, 2015).

More recently, an increasing number of sensor and other measurement devices have been connected to the internet, giving rise to the so-called Internet of Things. It also includes data collected through participatory, crowdsourcing or citizen science mechanisms (Heipke, 2010; Kamel Boulos et al., 2011; Chunara et al., 2013). The opportunities and challenges arising from the Internet of Things are only just being recognised by manufacturing industries, and this has been referred to as the fourth industrial revolution or *Industry 4.0* (Lee et al., 2014).

In animal production, precision livestock farming is considered to have significant potential to improve animal health, production and welfare. While sensor technology is already used, for example, in dairy cattle feeding, mastitis, fertility, locomotion and metabolism, the integration and analysis of the data for decision making still needs further development (Rutten et al., 2013; Mortari and Lorenzelli, 2014). It is very likely that more widespread utilisation and better adaptation of these digital technologies will provide an opportunity for more effective traceability of livestock and their products and animal health surveillance. However, to get the most out of both Big Data and data generated by the Internet of Things requires a change in analytical approach, which has led to the development of data science.

3. Data science

While the amount of data available for analysis continues to increase exponentially, the development of suitable analytical tools for converting this raw data into useful knowledge has been much slower (Anon., 2013; Kambatla et al., 2014; Gandomi and Haider, 2015). Up until about five to ten years ago, the challenge associated with analysing epidemiological data sourced from existing, and sometimes collated across, multiple data sources was addressed using secondary data analysis approaches (Sorensen et al., 1996; Olsen, 2008). The associated methodological developments were strongly underpinned by the well-established principles of the scientific method, with data analysis primarily being the responsibility of statistical science. While most epidemiologists will have experienced the technical challenges associated with data management, it is notable that most postgraduate epidemiology training programmes primarily focus on tabular, while barely covering, relational databases.

With Big Data, we have now reached levels of data complexity as expressed in the first four of the 5 V attributes (i.e. volume, velocity, variety, veracity and value) which cannot be effectively addressed by the 'classical' data management and analytical skill set. As result, and strongly incentivised by businesses interested in exploiting value, the fifth of the Big Data attributes, data analysts with advanced computer science skills have now become involved, in order to effectively convert the variety of data types and sources into knowledge (Wing, 2008; Bell et al., 2009; Porter et al., 2012).

An extreme interpretation of this new situation was expressed by the Editor-in-Chief of *Wired Magazine* in an article entitled "*The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*" (Anderson, 2008). He suggested that in the Petabyte Age, hypothesis-driven research would become irrelevant and be replaced by mining of data for associations. This extreme view

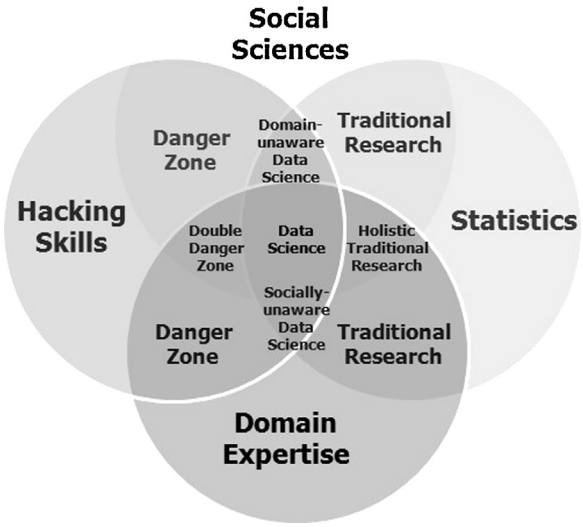

Fig. 1. Four-bubble Data Science Venn diagram (reproduced with permission from Malak, 2014).

has resulted in some debate (Norvig, 2009; Pigliucci, 2009; Schutt and O'Neil, 2013; Faghmous and Kumar, 2014; Mayer-Schönberger and Cukier, 2014). Arguably, this debate has had the benefit that scientists reflected on the utility of their respective research methodologies. It has also re-emphasised the importance of knowledge discovery going beyond the development of descriptive models optimised for mathematical accuracy. The renewed belief in the development of increasingly more powerful artificial intelligence applications also emphasises the effectiveness of machine learning algorithms (Jones, 2014; Gibney, 2015; Scholkopf, 2015; You, 2015). This does not mean that statistical algorithms will become less relevant, as is reflected in the current trend towards synergistic method development between the computer and statistical sciences (Kuhn and Johnson, 2014; Peters et al., 2014).

To more effectively deal with Big Data, and the associated analysis challenges, the new discipline of *data science* has been established which explicitly requires a multidisciplinary team approach (Dhar, 2013; Schutt and O'Neil, 2013). The four-bubble Data Science Venn diagram (Fig. 1) adapted from the three-bubble original by Conway (2010) reflects the interdependence between required disciplines (Malak, 2014). As such, it emphasises the importance of integrating computer science, statistical science, specialist domain expertise and social science. Conway (2010) had not explicitly separated social science from specialist domain expertise, but it seems justified to separate it out given that human behaviour has a major influence on the characteristics of most data sources. Arguably, this perspective is very similar to the interdisciplinary approach that underpins One Health and Ecohealth.

Gartner Inc. (Gartner, 2014), an international information technology research and advisory company, annually evaluates the maturity of emerging technologies and presents their conclusions using the *Gartner Hype Cycle* (Fig. 2). By representing time on the *x*-axis and expectations on the *y*-axis, they define five phases through which a technology will typically pass before it potentially achieves widespread adoption. Starting with the Innovation Trigger phase and rapidly climbing the Peak of Inflated Expectations, the cycle then descends into the Trough of Disillusionment (with respect to expectations). From there it may ascend the Slope of Enlightenment before finally reaching the Plateau of Productivity. As of 2014, the *Gartner Hype Cycle* considered data science (entering the Peak of Inflated Expectations) to be lagging behind both the Internet of Things (midway through the Peak) and Big Data (entering the Trough of Disillusionment) (Gartner, 2014) – a trend that mirrors

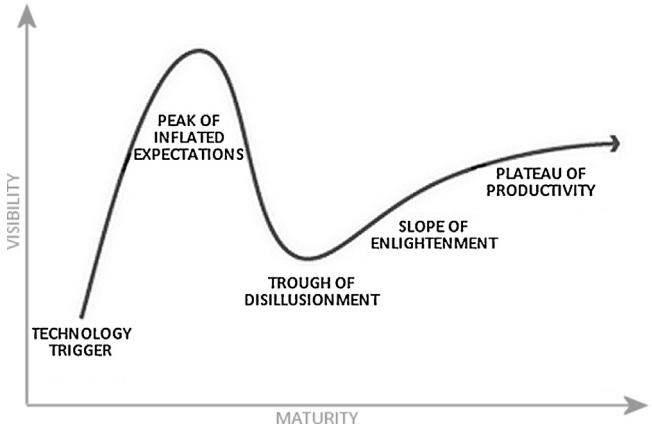

Fig. 2. The generic Gartner Hype Cycle which defines the five phases through which a technology will typically pass before it potentially achieves widespread adoption (reproduced with permission from Gartner, 2014; Gartner Methodologies, Hype Cycle, http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp).

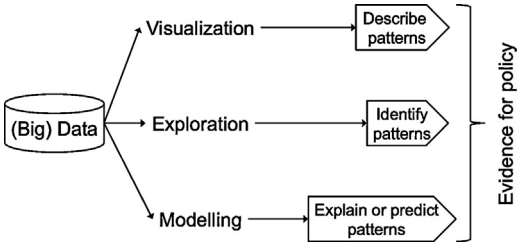

Fig. 3. Spatial and temporal data analysis in support of decision making in animal health in the Big Data era.

the development of spatial analytical methods suitable for taking advantage of the opportunities offered by georeferenced Big Data.

4. The development of spatial analytical methods

The analysis framework based on Pfeiffer et al. (2008a), presented in a slightly updated format in Fig. 3, is still relevant for structuring the different spatial and spatio-temporal epidemiological analytical methods. These are based primarily on classical statistical theory, with the addition of Bayesian methods to address the issue of spatial and temporal dependence. However, the assumptions, in particular, of frequentist statistical methods are usually not met for Big Data, and therefore analytical algorithms are required which are statistically robust (i.e. non-parametric) and also are capable of efficiently analysing very large datasets. The developments for epidemiological analyses have, so far, been primarily through the inclusion of machine learning regression amongst the modelling methods, whereas in visualisation and exploration it has been primarily through more effective use of interfaces and flexible software environments. It needs to be emphasised that data analysis plays an important role also in dealing with the five Vs of Big Data (i.e. volume, velocity, variety, veracity and value). The first four attributes refer to areas which are subject to significant research aimed at optimisation of data utilisation. A particularly difficult aspect is the challenge presented by differences in data quality, including the ubiquitous presence and heterogeneity of bias. Below, we discuss developments for each of the three analysis categories of the framework.

4.1. Visualisation of spatial patterns

Visualisation, whether as part of the analysis process or communication purposes, has always been a particular strength of spatial analysis and so it is not surprising that the biggest advances in the field of spatial analysis, with respect to Big Data, have occurred in this area. Big Data analytics emphasise the use of interactive visualisation methods using charts and maps, so that analysts and decision makers can quickly obtain insights from the most up-to-date data (e.g. GAPMINDER; http://www.gapminder.org, last accessed May 2015).

While geographical information system (GIS) software remains at the forefront for manipulating and producing complex visualisations of spatio-temporal data, the advent of interactive digital maps and virtual globes such as Google Maps™ and Google Earth™ has encouraged simple visualisation of disease data in real time, as illustrated by the integration of such digital platforms into an ever-expanding number of animal and public-health projects and platforms. For example, HealthMap (http://www.healthmap.org, last accessed April 2015), together with its mobile app *Outbreaks Near Me* (http://www.healthmap.org/outbreaksnearme/, last accessed May 2015), provides real-time surveillance of emerging public health threats (Brownstein et al., 2008; Freifeld et al., 2008), while *Nature*'s use of the platform to track the global spatio-temporal spread of highly pathogenic avian influenza H5N1 (Google Earth Avian Flu; http://www.nature.com/nature/multimedia/googleearth, last accessed April 2015; Butler, 2006) won the Association of Online Publishers (AOP) Use of a New Digital Platform Award in 2006.

Google Earth has also proved valuable for visualising disease data from informal settlements or rural areas in developing countries where the lack of geolocation infrastructure such as road names or house numbers precludes the use of conventional mapping software for visualising disease data. In a modern day reprise of John Snow's 1856 cholera investigation, use of the digital platform allowed Baker et al. (2011) to map the spread of a typhoid outbreak in Kathmandu – where street names are not used – and trace the cause of the epidemic to low-lying public water resources.

In addition to web-based mapping of disease, a related field is that of volunteered geographic information (VGI) (Goodchild, 2007; Goodchild and Li, 2012) or crowdsourced cartography (Dodge and Kitchin, 2013) which uses volunteers to create maps. A well-known example of VGI is OpenStreetMap (OSM; http://www.openstreetmap.org/, last accessed May 2015), an open, online, editable map of the world being created by volunteers using a combination of local knowledge, GPS tracks and aerial imagery. During the 2014 West Africa Ebola crisis, personnel of Médecins Sans Frontières (MSF) enlisted the help of the Humanitarian OSM Team (HOT) – an extension of OSM – to map Guéckédou – the main city in Guinea affected by the outbreak (Hodson, 2014). Within 20 h of receiving the request, online volunteers had mapped three cities in Guinea based on satellite imagery of the area, populating them with over 100 000 buildings – information that proved crucial for door-to-door canvassing of inhabitants and mapping the spread of disease. Other examples of crowdsourced cartography include GeoWiki (http://www.geo-wiki.org/, last accessed May 2015), a global network of volunteers working to improve the quality of global land-cover maps.

In a systematic review of visualisation and analytics for infectious disease research, Carroll et al. (2014) identified limitations of visualisation tools in terms of their utility and usability for end users, including risk of misinterpretation of choropleth maps by not adequately showing missing data and uncertainty. They report a need for interdisciplinary tool development to allow valid integrated analysis of data sourced from different areas such as molecular, network and population data. Similarly, not all crowd-

sourced information is of equal quality; some data are of higher quality than others just as some contributors are consistently better than others (Haklay, 2010). The inclusion of robust measures of quality for VGI would be useful to indicate the level of confidence associated with each piece of information, and although traditional statistical concepts of uncertainty and bias are hard to apply to VGI, other options are available. For example, See et al. (2013) found that when classifying land-cover, volunteer accuracy appeared to be higher when responses for a given location were more consistent and when the volunteers indicated higher confidence in their responses, suggesting that these additional pieces of information could be used to develop associated robust measures of quality. Additional possibilities include the application of Bayesian probability or Dempster–Shafer theory (Eastman, 2009) to provide measures of confidence.

Another area that has received significant attention is the analysis of molecular, movement and network data (Brunker et al., 2012; Okabe and Sugihara, 2012; Andrienko and Andrienko, 2013; Carrel and Emch, 2013). In this context, the utility of mobile phone call location records for infectious disease research and policy development has been of recent interest (Tatem, 2014; Wesolowski et al., 2014b). For example, mobile call location records were used during the 2014 Ebola outbreak to visualise and quantify the movements of a sample of the human population in West Africa (Wesolowski et al., 2014a), effectively visualising the spatial catchment areas of urban centres which reached even the more distant locations of the region.

4.2. Exploration of spatial data

Spatial exploratory analysis uses statistical methods to test the likelihood that an observed spatial or spatio-temporal pattern is a result of chance variation. Amongst these, the spatial and space-time scan statistic are probably the most often used cluster detection methods. In recent years, the scan statistic has been further developed to incorporate diverse spatial structures and a range of outcome variables with different measurement scales (Correa et al., 2014; Costa and Kulldorff, 2014; Murray et al., 2014; Prates et al., 2014).

Similarly, interpolation methods for spatial data, such as kriging, have also been expanded to accommodate different types of outcome variables such as ordinal or Poisson measurement scales (Li and Heap, 2014; Oliver and Webster, 2014). However, kernel smoothing – used to convert point data into smooth raster maps and an effective tool for visualising continuous spatial variation in risk and rates – still requires continuing methodological development, particularly in the selection of appropriate bandwidths for kernel functions (Sarojinie Fernando and Hazelton, 2014).

4.3. Spatial modelling

Modelling approaches can be broadly categorised into data- and knowledge-driven methods (Pfeiffer et al., 2008b; Stevens and Pfeiffer, 2011). The former use a dataset comprising several risk factors together with an outcome variable, and risk-factor effect estimates are usually obtained using one of a range of regression methods. Data-driven approaches can be further sub-divided depending on whether they require both disease presence and absence data to calibrate the model, or presence-only data. Alternatively, with knowledge-driven methods, risk estimates are derived based on existing or hypothesised understanding of the causal relationships leading to disease occurrence (Stevens and Pfeiffer, 2011; Stevens et al., 2013).

Amongst presence-absence data-driven methods, Bayesian approaches used to be a major focus of development but these have recently been complemented by machine learning methods which are better able to deal with the large datasets of the Big Data era (Vatsavai et al., 2012; Lawson, 2014; Peters et al., 2014,b; van Zyl, 2014a,b; Ziegler and König, 2014). Machine learning regression modelling used to consist primarily of classification tree analysis (Breiman et al., 1984), but in recent years this approach has been more or less replaced by random forest and boosted regression tree methods. These approaches are considered to be less affected by missing values, non-linearity, autocorrelation, lack of independence and distributional assumptions than parametric methods. In addition, several comparative reviews of the performance of the different species distribution modelling methods (Hirzel et al., 2006; Elith and Graham, 2009; França and Cabral, 2015) suggest that, in general, tree-based regression methods tend to perform slightly better than other spatial regression approaches. Requiring large datasets to be able to produce generalizable inferences, these methods are ideally suited for analysing Big Data.

Boosted regression trees are being used with increasing frequency to predict species distributions and disease risk (Hay et al., 2006; Martin et al., 2011; Gilbert et al., 2014; Pigott et al., 2014), while Tatem et al. (2014) used random forest regression tree analysis to generate risk maps for malaria occurrence and human movement flows based on mobile phone call location records to describe the spatial variation in malaria exportation/importation potential for Namibia.

However, a common problem with disease regression modelling is that, while the outcome variable may consist of fairly reliable disease presence information, for a usually unknown number of space-time observations, absence of disease reporting may not reflect true absence of disease or absence data may not be available (e.g. surveillance data). This is also common in ecological species distribution modelling and has led to the development of different sampling approaches to generate pseudo-absence data that can be used with regression methods requiring both presence and absence data, as well as the development of specific modelling techniques requiring presence-only data such as the ecological niche modelling (ENM) methods including ecological niche factor analysis (ENFA), genetic algorithm for rule-set production (GARP) and maximum entropy (Maxent) (Hirzel et al., 2002; Elith and Leathwick, 2009). Requiring only disease presence data means that ENM methods can make use of the extensive disease occurrence data available in surveillance databases, and by extension, of web-based Big Data systems containing information on location of disease occurrence but lacking absence data.

Increased access to molecular information on hosts and pathogens has resulted in the emergence of the field of phylogeography which integrates geospatial with genetic data (Liang et al., 2010; Chan et al., 2011; Faria et al., 2011; Pybus et al., 2012; Carrel and Emch, 2013; Alvarado-Serrano and Knowles, 2014). Further, combining ENM and phylogeography can be particularly informative for studies of globally distributed pathogens where environmental associations may be linked to genetic variation. From a methodological perspective, knowing that certain lineages exhibit niche specialisation and unique geographic distributions, can improve model accuracy by dividing a large population into biologically meaningful sub-populations (Mullins et al., 2013). However, ignoring such genetic variations may result in ENM which are biased towards a dominant strain in a particular region. There are also now a number of examples of integrated analysis of spatial and social network data (Firestone et al., 2011; Giebultowicz et al., 2011; Firestone et al., 2012).

Hay et al. (2013) discussed the opportunities arising from taking advantage of Big Data through integrated analyses and emphasises the need for dynamic, risk-mapping capability based on integrated analysis ranging from more static environmental, to highly dynamic, social media risk factor variables.

While data-driven methods still dominate in spatial modelling, the use of knowledge-driven approaches has increased during the last ten years. This is particularly the case for dynamic modelling, but also for static approaches such as multi-criteria decision analysis (MCDA). A key characteristic of these modelling approaches is their emphasis on inter-disciplinarity in that system understanding generated by different disciplines needs to be integrated so that the particular modelling objectives can be meaningfully achieved. Big Data is unlikely to result in the demise of the need for use of expert opinion and integration of existing knowledge such as MCDA, particularly in the context of management of new and emerging risks.

Use of knowledge-driven approaches and interpretation of results needs to recognise the potential impact of bias and underestimation of variability, given that the model structure is based on the opinion of experts, and the parameters tend to also be based on expert opinion or generated by a variety of research activities. Malczewski (2006) in his review of spatial MCDA notes that the methodology has been applied in many areas, particularly for land suitability analysis, and that it facilitated the development of participatory GIS. However, he highlights that the methodologies are frequently used without taking account of the method's underlying assumptions. More recently, Malczewski (2010) and Hongoh et al. (2011) emphasised the benefits of using spatially explicit MCDA to improve transparency and trans-disciplinarity of decision-making processes.

In animal health, Clements et al. (2006) and Stevens et al. (2013) used spatial MCDA to generate suitability maps for Rift Valley fever for Africa and avian influenza H5N1 for Asia, respectively. Both applied Dempster–Shafer theory to explicitly express and propagate uncertainty in relation to knowledge about the underlying processes expressed in the decision rules. de Glanville et al. (2014) generated suitability maps for African swine fever for Africa and used Monte-Carlo sensitivity analysis to express uncertainty in relation to model outputs. Other animal health applications of spatial MCDA have addressed animal diseases such as African horse sickness in Spain and Rift Valley fever in Italy (Tran et al., 2013; Sanchez-Matamoros et al., 2014). The increasing use of MCDA in the environmental sciences has resulted in further development of MCDA methodologies to reduce the influence of subjectivity of individual criteria weights on the risk score outcome (Yemshanov et al., 2013; Feizizadeh et al., 2014; Jankowski et al., 2014; Ligmann-Zielinska and Jankowski, 2014).

5. Conclusions

It is almost certain that in the near future humanity will have to deal with major infectious disease threats, largely as either a direct or indirect consequence of anthropogenic development. The latter involves technological and scientific advances which have, and will, generate opportunities for more effective management of current, and new and emerging infectious disease threats. Big Data, together with the Internet of Things, has introduced a new way of collecting and analysing data that is very different from the hypothesis-driven approaches previously accepted by the international scientific community as the primary mechanism for generating new scientific knowledge. Within the area of epidemiological analysis of spatial and spatio-temporal data, Big Data associated technologies and data sources have, so far, had limited impact, with the main advances having been associated with machine learning modelling methods, the recent use of mobile phone location records, molecular diagnostic and animal movement data. To more effectively harness the opportunities offered by these new digital technologies in animal and human health, an interdisciplinary approach will have to be embraced which, in addition to the various scientific domains associated with human, animal and environmental health, also includes computer science. This will result in a particularly interesting situation for epidemiologists whose scientific strength has been the integration between the applied health sciences and the more theoretical and abstract methods underpinning statistical analysis, to which they could now add the role of acting as an interface with the computer science aspects of Big Data and the Internet of Things. By doing so they will be able to continue their substantial contribution to the understanding of cause-effect relationships in eco-social systems, and thereby expand the knowledge-base underpinning effective animal health risk management.

Conflict of interest

The authors report no conflict of interest.

Acknowledgements

The first author would like to express his appreciation to Prof Roger Morris, the recipient of the 2014 Calvin Schwabe Lifetime Achievement award, for inspiring him to pursue a career in veterinary epidemiology and for the scientific mentorship provided while working together over a period of 11 years.

References

Alvarado-Serrano, D.F., Knowles, L.L., 2013. Ecological niche models in phylogeographic studies: applications, advances and precautions. Mol. Ecol. Resources 14, 233–248.
Anderson, C., 2008. The end of theory: the data deluge makes the scientific method obsolete. Wired Mag. 16, 07.
Andrienko, N., Andrienko, G., 2012. Visual analytics of movement: an overview of methods, tools and procedures. Inf. Visual. 12, 3–24.
Anon, 2011. Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease. The National Academies Press, Washington DC, USA.
Anon, 2013. Frontiers in Massive Data Analysis. National Research Council of the National Academies, Washington DC, USA, 176pp.
Anon, 2014a. Big Data and Privacy A Technological Perspective. Executive Office of the President, President's Council of Advisors on Science and Technology, Washington DC, USA, 57pp.
Anon, 2014b. The Internet of Things: Making the Most of the Second Digital Revolution. The Government Office for Science, London, UK, 38pp.
Baker, S., Holt, K.E., Clements, A.C.A., Karkey, A., Arjyal, A., Boni, M.F., Dongol, S., Hammond, N., Koirala, S., Duy, P.T., Nga, T.V.T., Campbell, J.I., Dolecek, C., Basnyat, B., Dougan, G., Farrar, J.J., 2011. Combined high-resolution genotyping and geospatial analysis reveals modes of endemic urban typhoid fever transmission. Open Biol. 1, 110008, http://dx.doi.org/10.1098/rsob.110008
Bell, G., Hey, T., Szalay, A., 2009. Computer science. Beyond the data deluge. Science 323, 1297–1298.
Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth & Brooks Monterey, California, pp. USA–$9.
Brownstein, J.S., Freifeld, C.C., Madoff, L.C., 2009. Digital disease detection – harnessing the web for public health surveillance. N. Engl. J. Med. 360, 2153–2157.
Brownstein, J.S., Freifeld, C.C., Reis, B.Y., Mandl, K.D., 2008. Surveillance sans frontieres: internet-based emerging infectious disease intelligence and the HealthMap project. PLoS Med. 5, e151.
Brownstein, J.S., Freifeld, C.C., Chan, E.H., Keller, M., Sonricker, A.L., Mekaru, S.R., Buckeridge, D.L., 2010. Information technology and global surveillance of cases of 2009H1N1 influenza. N. Engl. J. Med. 362, 1731–1735.
Brunker, K., Hampson, K., Horton, D.L., Biek, R., 2012. Integrating the landscape epidemiology and genetics of RNA viruses: rabies in domestic dogs as a model. Parasitology 139, 1899–1913.
Butler, D., 2006. Mashups mix data into global service. Nature 439, 6–7.
Butler, D., 2013. When Google got flu wrong. Nature 494, 155–156.
Carneiro, H.A., Mylonakis, E., 2009. Google trends: a web-based tool for real-time surveillance of disease outbreaks. Clin. Infect. Dis. 49, 1557–1564.
Carrel, M., Emch, M., 2013. Genetics: a new landscape for medical geography. Ann. Assoc. Am. Geogr. 103, 1452–1467.
Carroll, L.N., Au, A.P., Detwiler, L.T., Fu, T.-C., Painter, I.S., Abernethy, N.F., 2014. Visualization and analytics tools for infectious disease epidemiology: a systematic review. J. Biomed. Inf. 51, 287–298.
Chan, L.M., Brown, J.L., Yoder, A.D., 2011. Integrating statistical genetic and geospatial methods brings new power to phylogeography. Mol. Phylogenet. Evol. 59, 523–537.
Chew, C., Eysenbach, G., 2010. Pandemics in the age of Twitter: content analysis of tweets during the 2009H1N1 outbreak. PLoS ONE 5, e14118.

Chunara, R., Smolinski, M., Brownstein, J., 2013. Why we need crowdsourced data in infectious disease surveillance. Curr. Infect. Dis. Rep. 15, 316–319.
Clements, A.C.A., Pfeiffer, D.U., Martin, V., 2006. Application of knowledge-driven spatial modelling approaches and uncertainty management to a study of Rift Valley fever in Africa. Int. J. Health Geographics 5, 57.
Collier, N., Doan, S., Kawazoe, A., Goodwin, R.M., Conway, M., Tateno, Y., Ngo, Q.-H., Dien, D., Kawtrakul, A., Takeuchi, K., Shigematsu, M., Taniguchi, K., 2008. BioCaster: detecting public health rumors with a web-based text mining system. Bioinformatics 24, 2940–2941.
Collins, F.S., Varmus, H., 2015. A new initiative on precision medicine. N. Engl. J. Med. 372, 793–795.
Conway, D., 2010. The data science Venn Diagram. (http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram, last accessed 04.15.).
Correa, T.R., Assuncao, R.M., Costa, M.A., 2014. A critical look at prospective surveillance using a scan statistic. Stat. Med. 34, 1081–1093.
Costa, M., Kulldorff, M., 2014. Maximum linkage space-time permutation scan statistics for disease outbreak detection. Int. J. Health Geographics 13, 20, http://dx.doi.org/10.1186/1476-072X-13-20
Cowen, P., Garland, T., Hugh-Jones, M.E., Shimshony, A., Handysides, S., Kaye, D., Madoff, L.C., Pollack, M.P., Woodall, J., 2006. Evaluation of ProMED-mail as an electronic early warning system for emerging animal diseases: 1996 to 2004. J. Am. Vet. Med. Assoc. 229, 1090–1099.
de Glanville, W.A., Vial, L., Costard, S., Wieland, B., Pfeiffer, D.U., 2014. Spatial multi-criteria decision analysis to predict suitability for African swine fever endemicity in Africa. BMC Vet. Res. 10, 9, http://dx.doi.org/10.1186/1746-6148-10-9
Dhar, V., 2013. Data science and prediction. Commun. ACM 56, 64–73.
Dodge, M., Kitchin, R., 2013. Crowdsourced cartography: mapping experience and knowledge. Environ. Plann. A 45, 19–36.
Eastman, J.R., 2009. Decision support: uncertainty management. In: IDRISI Guide to GIS and Image Processing, Accessed in IDRISI Andes, Worcester, MA: Clark University, pp. 156–172.
Elith, J., Graham, C.H., 2009. Do they? How do they? Why do they differ? On finding reasons for differing performances of species distribution models. Ecography 32, 66–77.
Elith, J., Leathwick, J.R., 2009. Species distribution models: ecological explanation and prediction across space and time. Annu. Rev. Ecol. Evol. Syst. 40, 677–697.
Eysenbach, G., 2001. What is e-health. J. Med. Internet Res. 3, e20.
Faghmous, J.H., Kumar, V., 2014. A big data guide to understanding climate change: the case for theory-guided data science. Big Data 2, 155–163.
Faria, N.R., Suchard, M.A., Rambaut, A., Lemey, P., 2011. Toward a quantitative understanding of viral phylogeography. Curr. Opin. Virol. 1, 423–429.
Feizizadeh, B., Jankowski, P., Blaschke, T., 2014. A GIS-based spatially-explicit sensitivity and uncertainty analysis approach for multi-criteria decision analysis. Comput. Geosci. 64, 81–95.
Fernández, A., del Río, S., López, V., Bawakid, A., del Jesus, M.J., Benítez, J.M., Herrera, F., 2014. Big data with cloud computing: an insight on the computing environment, MapReduce, and programming frameworks. Wiley Interdiscip. Rev.: Data Min. Knowl. Discovery 4, 380–409.
Firestone, S.M., Christley, R.M., Ward, M.P., Dhand, N.K., 2012. Adding the spatial dimension to the social network analysis of an epidemic: investigation of the 2007 outbreak of equine influenza in Australia. Prev. Vet. Med. 106, 123–135.
Firestone, S.M., Ward, M.P., Christley, R.M., Dhand, N.K., 2011. The importance of location in contact networks: describing early epidemic spread using spatial social network analysis. Prev. Vet. Med. 102, 185–195.
França, S., Cabral, H.N., 2015. Predicting fish species richness in estuaries: which modelling technique to use. Environ. Model. Software 66, 17–26.
Freifeld, C.C., Mandl, K.D., Reis, B.Y., Brownstein, J.S., 2008. HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. J. Am. Med. Inf. Assoc. 15, 150–157.
Gandomi, A., Haider, M., 2015. Beyond the hype: big data concepts, methods, and analytics. Int. J. Inf. Manage. 35, 137–144.
Gartner, 2014. Gartner's 2014Hype Cycle for Emerging Technologies Maps the Journey to Digital Business. Gartner, Inc., pp. 2014.
Gibney, E., 2015. DeepMind algorithm beats people at classic video games. Nature 518, 465–466.
Giebultowicz, S., Ali, M., Yunus, M., Emch, M., 2011. The simultaneous effects of spatial and social networks on cholera transmission. Interdiscip. Perspect. Infect. Dis. 2011, http://dx.doi.org/10.1155/2011/604372, Article ID 604372.
Gilbert, M., Golding, N., Zhou, H., Wint, G.R., Robinson, T.P., Tatem, A.J., Lai, S., Zhou, S., Jiang, H., Guo, D., Huang, Z., Messina, J.P., Xiao, X., Linard, C., Van Boeckel, T.P., Martin, V., Bhatt, S., Gething, P.W., Farrar, J.J., Hay, S.I., Yu, H., 2014. Predicting the risk of avian influenza A H7N9 infection in live-poultry markets across Asia. Nat. Commun. 5, http://dx.doi.org/10.1038/ncomms5116, Article number 4116.
Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L., 2008. Detecting influenza epidemics using search engine query data. Nature 457, 1012–1014.
Goodchild, M., 2007. Citizens as sensors: the world of volunteered geography. GeoJournal 69, 211–221.
Goodchild, M.F., Li, L., 2012. Assuring the quality of volunteered geographic information. Spat. Stat. 1, 110–120.
Grein, T.W., Kamara, K.B., Rodier, G., Plant, A.J., Bovier, P., Ryan, M.J., Ohyama, T., Heymann, D.L., 2000. Rumors of disease in the global village: outbreak verification. Emerg. Infect. Dis. 6, 97–102.
Haklay, M., 2010. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. Environ. Plann. B: Plann. Des. 37, 682–703.
Hartley, D.M., Nelson, N.P., Arthur, R.R., Barboza, P., Collier, N., Lightfoot, N., Linge, J.P., van der Goot, E., Mawudeku, A., Madoff, L.C., Vaillant, L., Walters, R., Yangarber, R., Mantero, J., Corley, C.D., Brownstein, J.S., 2013. An overview of Internet biosurveillance. Clin. Microbiol. Infect. 19, 1006–1013.
Hay, S.I., Graham, A., Rogers, D.J., 2006. Global mapping of infectious diseases: methods, examples and emerging applications. Emerg. Infect. Dis. 13, 674, http://dx.doi.org/10.3201/eid1304.070037
Hay, S.I., George, D.B., Moyes, C.L., Brownstein, J.S., 2013. Big data opportunities for global infectious disease surveillance. PLoS Med. 10, e1001413, http://dx.doi.org/10.1371/journal.pmed.1001413
Heipke, C., 2010. Crowdsourcing geospatial data. ISPRS J. Photogramm. Remote Sens. 65, 550–557.
Heymann, D.L., Rodier, G.R., 2001. Hot spots in a wired world: WHO surveillance of emerging and re-emerging infectious diseases. Lancet Infect. Dis. 1, 345–353.
Hirzel, A.H., Hausser, J., Chessel, D., Perrin, N., 2002. Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data. Ecology 83, 2027–2036.
Hirzel, A.H., Le Lay, G., Helfer, V., Randin, C., Guisan, A., 2006. Evaluating the ability of habitat suitability models to predict species presences. Ecol. Model. 199, 142–152.
Hodson, H., 2014. Online army helps map Guinea's Ebola outbreak. New Sci. 2964, (http://www.newscientist.com/article/mg22229644.400-online-army-helps-map-guineas-ebola-outbreak.html#.VToUHfJ0yM8; last accessed 04.15).
Hongoh, V., Hoen, A.G., Aenishaenslin, C., Waaub, J.P., Belanger, D., Michel, P., Lyme, M.C., 2011. Spatially explicit multi-criteria decision analysis for managing vector-borne diseases. Int. J. Health Geographics 10, 70, http://dx.doi.org/10.1186/1476-072X-10-70
Istepanian, R., Jovanov, E., Zhang, Y., 2004. Introduction to the special section on m-Health: beyond seamless mobility and global wireless health-care connectivity. IEEE Trans. Inf. Technol. Biomed. 8, 405–414.
Jankowski, P., Fraley, G., Pebesma, E., 2014. An exploratory approach to spatial decision support. Comput. Environ. Urban Syst. 45, 101–113.
Jones, N., 2014. Computer science: the learning machines. Nature 505, 146–148.
Kambatla, K., Kollias, G., Kumar, V., Grama, A., 2014. Trends in big data analytics. J. Parallel Distrib. Comput. 74, 2561–2573.
Kamel Boulos, M., Al-Shorbaji, N., 2014. On the Internet of Things, smart cities and the WHO Healthy Cities. Int. J. Health Geographics 13, 10.
Kamel Boulos, M., Resch, B., Crowley, D., Breslin, J., Sohn, G., Burtner, R., Pike, W., Jezierski, E., Chuang, K.-Y., 2011. Crowdsourcing, citizen sensing and sensor web technologies for public and environmental health surveillance and crisis management: trends, OGC standards and application examples. Int. J. Health Geographics 10, 67, http://dx.doi.org/10.1186/1476-072X-10-67
Keller, M., Blench, M., Tolentino, H., Freifeld, C., Mandl, K., Mawudeku, A., Eysenbach, G., Brownstein, J., 2009. Use of unstructured event-based reports for global infectious disease surveillance. Emerg. Infect. Dis. 15, 689–695.
Kuhn, M., Johnson, K., 2014. Who's afraid of the big black box? Statisticians' vital role in big data and predictive modelling. Significance 11, 35–37.
Lawson, A.B., 2014. Hierarchical modeling in spatial epidemiology. Wiley Interdiscip. Rev.: Comput. Stat. 6, 405–417.
Lazer, D., Kennedy, R., King, G., Vespignani, A., Big data, 2014. The parable of Google Flu: traps in big data analysis. Science 343, 1203–1205.
Lee, J., Kao, H.-A., Yang, S., 2014. Service innovation and smart analytics for Industry 4. 0 and Big Data environment. Procedia CIRP 16, 3–8.
Leetaru, K., 2014. Why big data missed the early warning signs of Ebola. Foreign Policy http://foreignpolicy.com/2014/09/26/why-big-data-missed-the-early-warning-signs-of-ebola/
Li, J., Heap, A.D., 2014. Spatial interpolation methods applied in the environmental sciences: a review. Environ. Model. Software 53, 173–189.
Liang, L., Xu, Chen, B., Liu, Y., Cao, Y., Fang, W., Feng, L., Goodchild, L., Gong, M.F., P, 2010. Combining spatial-temporal and phylogenetic analysis approaches for improved understanding on global H5N1 transmission. PloS ONE 5, e13575, http://dx.doi.org/10.1371/journal.pone.0013575
Ligmann-Zielinska, A., Jankowski, P., 2014. Spatially-explicit integrated uncertainty and sensitivity analysis of criteria weights in multicriteria land suitability evaluation. Environ. Model. Software 57, 235–247.
Malak, M., 2014. The fourth bubble in the Data Science Venn Diagram: social sciences (http://datascienceassn.org/content/fourth-bubble-data-science-venn-diagram-social-sciences; last accessed 04.15.).
Malczewski, J., 2006. GIS-based multicriteria decision analysis: a survey of the literature. Int, J, Geog. Inf. Sci. 20, 703–726.
Malczewski, J., 2010. Multiple criteria decision analysis and geographic information systems. In: Ehrgott, M., Greco, S., Figueira, J.R. (Eds.), Trends in Multiple Criteria Decision Analysis. Springer, New York, pp. 369–395.
Martin, V., Pfeiffer, D.U., Zhou, X., Xiao, X., Prosser, D.J., Guo, F., Gilbert, M., 2011. Spatial distribution and risk factors of highly pathogenic avian influenza (HPAI) H5N1 in China. PLoS Pathog. 7, e1001308.
Mayer-Schönberger, V., Cukier, K., 2014. Big Data: A Revolution That Will Transform How We Live, Work, and Think. Mariner Books, Houghton Mifflin Harcourt Boston, 272pp.
Milinovich, G.J., Magalhães, R.J.S., Hu, W., 2015. Role of big data in the early detection of Ebola and other emerging infectious diseases. Lancet Global Health 3, e20–e21.

Milinovich, G.J., Williams, G.M., Clements, A.C.A., Hu, W., 2014. Internet-based surveillance systems for monitoring emerging infectious diseases. Lancet Infect. Dis. 14, 160–168.
Mortari, A., Lorenzelli, L., 2014. Recent sensing technologies for pathogen detection in milk: a review. Biosens. Bioelectron. 60, 8–21.
Mullins, J.C., Garofolo, G., Van Ert, M., Fasanella, A., Lukhnova, L., Hugh-Jones, M.E., Blackburn, J.K., 2013. Ecological niche modeling of Bacillus anthracis on three continents: evidence for genetic-ecological divergence. PLoS ONE 8, e72451.
Murray, A.T., Grubesic, T.H., Wei, R., 2014. Spatially significant cluster detection. Spat. Stat. 10, 103–116.
Mykhalovskiy, E., Weir, L., 2011. The Global Public Health Intelligence Network and early warning outbreak detection: a Canadian contribution to global public health. Can. J. Public Health 97, 42–44.
Norvig, P., 2009. All we want are the facts, ma'am (http://norvig.com/fact-check.html; last accessed 04.15.).
O'Driscoll, A., Daugelaite, J., Sleator, R.D., 2013. 'Big data', Hadoop and cloud computing in genomics. J. Biomed. Inf. 46, 774–781.
Okabe, A., Sugihara, K., 2012. Spatial Analysis Along Networks – Statistical and Computational Methods. John Wiley & Sons Chichester, UK, 306pp.
Oliver, M.A., Webster, R., 2014. A tutorial guide to geostatistics: computing and modelling variograms and kriging. CATENA 113, 56–69.
Olsen, J., 2008. Using secondary data. In: Rothman, K.J., Greenland, S., Lash, T.L. (Eds.), Modern Epidemiology. Lippincott Williams & Wilkins, Philadelphia, pp. 481–491.
Peters, D.P.C., Havstad, K.M., Cushing, J., Tweedie, C., Fuentes, O., Villanueva-Rosales, N., 2014. Harnessing the power of big data: infusing the scientific method with machine learning to transform ecology. Ecosphere 5, art67.
Pfeiffer, D.U., Robinson, T.P., Stevenson, M., Stevens, K.B., Rogers, D.J., Clements, A.C.A., 2008a. Introduction. In: Spatial Analysis in Epidemiology. Oxford University Press, Oxford, UK, pp. 1–8.
Pfeiffer, D.U., Robinson, T.P., Stevenson, M., Stevens, K.B., Rogers, D.J., Clements, A.C.A., 2008b. Spatial risk assessment and management of disease. In: Spatial Analysis in Epidemiology. Oxford University Press, Oxford, UK, pp. 119–120.
Pigliucci, M., 2009. The end of theory in science. EMBO Rep. 10, 534, http://dx.doi.org/10.1038/embor.2009.111
Pigott, D.M., Golding, N., Mylne, A., Huang, Z., Henry, A.J., Weiss, D.J., Brady, O.J., Kraemer, M.U., Smith, D.L., Moyes, C.L., Bhatt, S., Gething, P.W., Horby, P.W., Bogoch, I.I., Brownstein, J.S., Mekaru, S.R., Tatem, A.J., Khan, K., Hay, S.I., 2014. Mapping the zoonotic niche of Ebola virus disease in Africa. eLife 3, e04395.
Porter, J.H., Hanson, P.C., Lin, C.C., 2012. Staying afloat in the sensor data deluge. Trends Ecol. Evol. 27, 121–129.
Prates, M.O., Kulldorff, M., Assuncao, R.M., 2014. Relative risk estimates from spatial and space-time scan statistics: are they biased. Stat. Med. 33, 2634–2644.
Pybus, O.G., Suchard, M.A., Lemey, P., Bernardin, F.J., Rambaut, A., Crawford, F.W., Gray, R.R., Arinaminpathy, N., Stramer, S.L., Busch, M.P., Delwart, E.L., 2012. Unifying the spatial epidemiology and molecular evolution of emerging epidemics. Proc. Natl. Acad. Sci. U. S. A. 109, 15066–15071.
Rutten, C.J., Velthuis, A.G., Steeneveld, W., Hogeveen, H., 2013. Invited review: sensors to support health management on dairy farms. J. Dairy Sci. 96, 1928–1952.
Sanchez-Matamoros, A., Sanchez-Vizcaino, J.M., Rodriguez-Prieto, V., Iglesias, E., Martinez-Lopez, B., 2014. Identification of suitable areas for African horse sickness virus infections in Spanish equine populations. Transbound. Emerg. Dis., http://dx.doi.org/10.1111/tbed.12302
Sarojinie Fernando, W.T.P., Hazelton, M.L., 2014. Generalizing the spatial relative risk function. Spat. Spatio-Temporal Epidemiol. 8, 1–10.
Schadt, E.E., 2012. The changing privacy landscape in the era of big data. Mol. Syst. Biol. 8, 612.
Scholkopf, B., 2015. Artificial intelligence: learning to see and act. Nature 518, 486–487.
Schutt, R., O'Neil, C., 2013. Doing Data Science. O'Reilly Media Sebastopol, California, USA, 408pp.
See, L., Comber, A., Salk, C., Fritz, S., van der Velde, M., Perger, C., Schill, C., McCallum, I., Kraxner, F., Obersteiner, M., 2013. Comparing the quality of crowdsourced data contributed by expert and non-experts. PLoS ONE 8, e69958.
Signorini, A., Segre, A.M., Polgreen, P.M., 2011. The use of Twitter to track levels of disease activity and public concern in the U. S. during the Influenza A H1N1 pandemic. PLoS ONE 6, e19467.
Solanas, A., Patsakis, C., Conti, M., Vlachos, L.S., Ramos, V., Falcone, F., Postolache, O., Pérez-Martínez, P.A., Di Pietro, R., Perrea, D.N., Martínez-Ballesté, A., 2014. Smart health: a context-aware health paradigm within smart cities. IEEE Commun. Mag. 52, 74–81.
Sorensen, H.T., Sabroe, S., Olsen, J., 1996. A framework for evaluation of secondary data sources for epidemiological research. Int J Epidemiol 25, 435–442.
St Louis, C., Zorlu, G., 2012. Can Twitter predict disease outbreaks. Br. Med. J. 344, e2353, http://dx.doi.org/10.1136/bmj.e2353
Stevens, K.B., Pfeiffer, D.U., 2011. Spatial modelling of disease using data- and knowledge-driven approaches. Spat. Spatio-Temporal Epidemiol. 2, 125–133.
Stevens, K.B., Gilbert, M., Pfeiffer, D.U., 2013. Modeling habitat suitability for occurrence of highly pathogenic avian influenza virus H5N1 in domestic poultry in Asia: a spatial multicriteria decision analysis approach. Spat. Spatio-Temporal Epidemiol. 4, 1–14.
Tatem, A.J., 2014. Mapping population and pathogen movements. Int. Health 6, 5–11.
Tatem, A.J., Huang, Z., Narib, C., Kumar, U., Kandula, D., Pindolia, D.K., Smith, D.L., Cohen, J.M., Graupe, B., Uusiku, P., Lourenco, C., 2014. Integrating rapid risk mapping and mobile phone call record data for strategic malaria elimination planning. Malar. J. 13, 52, http://dx.doi.org/10.1186/1475-2875-13-52
Tolentino, H., Kamadjeu, R., Fontelo, P., Liu, F., Matters, M., Pollack, M., Madoff, L., 2007. Scanning the emerging infectious diseases horizon – visualizing ProMED emails using EpiSPIDER. Adv. Dis. Surveillance 2, 169.
Tran, A., Ippoliti, C., Balenghien, T., Conte, A., Gely, M., Calistri, P., Goffredo, M., Baldet, T., Chevalier, V., 2013. A geographical information system-based multicriteria evaluation to map areas at risk for Rift Valley fever vector-borne transmission in Italy. Transbound. Emerg. Dis. 60 (Suppl. 2), 14–23.
van Zyl, T., 2014a. Algorithmic considerations for geospatial and/or temporal big data. In: Karimi, H.A. (Ed.), Big Data– Techniques and Technologies in Geoinformatics. CRC Press, Boca Raton, Florida, USA, pp. 117–132.
van Zyl, T., 2014b. Machine learning on geospatial big data. In: Karimi, H.A. (Ed.), Big Data – Techniques and Technologies in Geoinformatics. CRC Press, Boca Raton, Florida, USA, pp. 133–148.
Vatsavai, R.R., Ganguly, A., Chandola, V., Stefanidis, A., Klasky, S., 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, ACM, Redondo Beach, California, pp. 1–10.
Wesolowski, A., Buckee, C.O., Bengtsson, L., Wetter, E., Lu, X., Tatem, A.J., 2014a. Commentary: containing the Ebola outbreak – the potential and challenge of mobile network data. PLoS Curr. Outbreaks 6, http://dx.doi.org/10.1371/currents.outbreaks.0177e7fcf52217b8b634376e2f3efc5e
Wesolowski, A., Stresman, G., Eagle, N., Stevenson, J., Owaga, C., Marube, E., Bousema, T., Drakeley, C., Cox, J., Buckee, C.O., 2014b. Quantifying travel behavior for infectious disease research: a comparison of data from surveys and mobile phones. Scientific Reports 4, 5678, http://dx.doi.org/10.1038/srep05678
Wilson, K., Brownstein, J.S., 2009. Early detection of disease outbreaks using the Internet. Can. Med. Assoc. J. 180, 829–831.
Wilson, K., von Tigerstrom, B., McDougall, C., 2008. Protecting global health security through the International Health Regulations: requirements and challenges. Can. Med. Assoc. J. 179, 44–48.
Wilson, N., Mason, K., Tobias, M., Peacey, M., Huang, Q.S., Baker, M., 2009. Interpreting Google Flu Trends data for pandemic H1N1 influenza: the New Zealand experience. Eurosurveillance, 14, pii=19386.
Wing, J.M., 2008. Computational thinking and thinking about computing. Philos. Trans. Ser. A. Math. Phys. Eng. Sci. 366, 3717–3725.
Yemshanov, D., Koch, F.H., Ben-Haim, Y., Downing, M., Sapio, F., Siltanen, M., 2013. A new multicriteria risk mapping approach based on a multiattribute frontier concept. Risk Anal. 33, 1694–1709.
You, J., 2015. Artifical intelligence: DARPA sets out to automate research. Science 347, 465.
Zeldenrust, M., Rahamat-Langendoen, J., Postma, M., van Vliet, J., 2008. The value of ProMED-mail for the Early Warning Committee in the Netherlands: more specific approach recommended. Eurosurveillance 13, 8033.
Ziegler, A., König, I.R., 2013. Mining data with random forests: current options for real-world applications. Wiley Interdiscip. Rev.: Data Min. Knowl. Discovery 4, 55–63.