Spatial and temporal data analysis in support of decision making for complex animal health problems in the Big Data era

Dirk U. Pfeiffer [a,*], Kim B. Stevens [a]

[a] Veterinary Epidemiology, Economics & Public Health Group, Dept. of Production & Population Health, Royal Veterinary College, London, UK

[*] Corresponding author at: Veterinary Epidemiology, Economics & Public Health Group, Dept. of Production & Population Health, Royal Veterinary College, Hawkshead Lane, Hatfield, Hertfordshire, AL97TA, United Kingdom. Tel.: +44 (1707) 666205; Fax: +44 (1707) 666574.
E-mail address: pfeiffer@rvc.ac.uk.

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 visualization 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 more widespread acceptance, also for spatial analysis, in the context of availability of large numbers of diverse data sources not suitable for integrated statistical analysis, published scientific information and the recognition for the need to use expert opinion to fill 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.

Keywords:

Big data; internet of things; data science; visualization; exploratory analysis; modelling, spatial analysis

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 rapidly increasing types and numbers of physical objects are connected through information networks. The so-called *Industry 4.0* 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*.

In the field of public health, the concepts of smart health, mHealth and eHealth can be seen as the starting point for these developments and, together with the recent increase in popularity of wearable sensors, have boosted the development of associated technologies. However, the 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, heterogenous data.

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

67 recent developments, and future directions, for spatial and temporal analysis in support of
68 managing complex animal health problems, starting with the different opportunities offered
69 by new data sources, followed by a discussion of the spatio-temporal approaches available for
70 analysing Big Data.

71 2. The data revolution: from the Internet via Big Data to the Internet of Things
72 Scientific approaches aimed at improving our understanding of the complexity of the systems
73 of which animal and human diseases form a part, usually involve data collection. However,
74 the way in which data are generated has changed radically over the last 30 years, mainly as a
75 result of the emergence of electronic methods for measuring, recording, storing and
76 distributing data. As part of this development, the Internet now forms the backbone of a
77 globally-reaching information network. The resulting Big Data has been embraced by the
78 business community but also represents an important opportunity for science.

79 Big data are generally characterized by 3Vs: volume (relative magnitude of dataset), velocity
80 (rate at which new data are generated) and variety (heterogenous structure of dataset [e.g.
81 text, video, audio]) (Gandomi and Haider, 2015). A fourth 'v' frequently used to describe
82 Big Data is veracity which acknowledges the inherent uncertainty frequently associated with,
83 in particular, web-based Big Data and the corresponding need for analytical approaches that
84 are able to account for this unreliability (Gandomi and Haider, 2015). Traditional database
85 management systems based on tabular or relational data management structures are not suited
86 to dealing with Big Data as most of it is unstructured. Cloud-based data storage using the
87 Apache Hadoop® distributed file system (http://hadoop.apache.org) has been developed to
88 allow efficient management of such data (O'Driscoll et al., 2013; Fernández et al., 2014).

89 A data mining approach was used to examine the frequency of particular words from a vast
90 number of digitised books published since the 1500s and their potential association with

historical events (Michel et al., 2011); for example, there was an association between the frequency of the word '*influenza*' and known historical occurrence of influenza epidemics. A similar methodology was used to explore the use of search term data for prediction of flu trends (Ginsberg et al., 2009a) 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/flutrends); 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 CDC Influenza Sentinel Provider Surveillance reports (Ginsberg et al., 2009b). 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 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 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 form an early-warning system in developing 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 (Osborne et al., 2001; Collier et al., 2006; Collier et al., 2008), EpiSPIDER (Tolentino et al., 2007; Keller et al., 2009), HealthMap (Osei-Bryson, 2003; Brownstein et al., 2008; Freifeld et al., 2008; Brownstein et al., 2009; Keller et al., 2009; Wilson and Brownstein, 2009; Brownstein et al., 2010), ProMED-mail (Ostle et al., 1986; 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 evidenced 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 9 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

141 Outbreak Alert and Response Network use such data as part of their day-to-day surveillance
142 activities (Grein et al., 2000; Heymann and Rodier, 2001) and are authorized to act on this
143 information (Wilson et al., 2008). Moving from surveillance to delivery of health care,
144 precision medicine aims to utilise Big Data for the purpose of optimising the use of
145 diagnostic tools, therapeutics and preventive management (Anonymous, 2011; Collins and
146 Varmus, 2015).

147 More recently, an increasing number of sensor and other measurement devices have become
148 connected to the internet. These have given rise to the so-called Internet of Things
149 (Anonymous, 2014b; Kamel Boulos and Al-Shorbaji, 2014). It also includes data collected
150 through participatory, crowdsourcing or citizen science mechanisms (Heipke, 2010; Kamel
151 Boulos et al., 2011; Chunara et al., 2013). The opportunities and challenges arising from the
152 Internet of Things are only just being recognised by manufacturing industries, and this has
153 been referred to as the fourth industrial revolution or Industry 4.0 (Lee et al., 2014). In animal
154 production, precision livestock farming is considered to have significant potential to improve
155 animal health, production and welfare. While sensor technology is already used, for example,
156 in dairy cattle feeding, mastitis, fertility, locomotion and metabolism, the integration and
157 analysis of the data for decision making still needs further development (Rutten et al., 2013;
158 Mortari and Lorenzelli, 2014). It is very likely that more widespread utilisation and better
159 adaptation of these digital technologies will provide an opportunity for more effective
160 traceability of livestock and their products and animal health surveillance. However, to
161 effectively use Big Data and that produced by the Internet of Things requires a change in
162 analytical approach which has led to the development of Data Science.

163 ## 3. Data Science

164 While the amount of data available for analysis continues to increase exponentially, the
165 development of suitable analytical tools for converting this raw data into useful knowledge

has been much slower (Anonymous, 2013; Kambatla et al., 2014; Gandomi and Haider, 2015). While statistical science has long been the discipline providing the primary skills and tools needed for data analysis, the inherent characteristics of Big Data mean that data analysts should now also have advanced computer science skills 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 - Will 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 has resulted in some debate (Norvig, 2009; Pigliucci, 2009; Schutt and O'Neil, 2013; Faghmous and Kumar, 2014; Mayer-Schönberger and Cukier, 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 adapted from the three-bubble original by Drew Conway reflects the interdependence between required disciplines (Malak, 2014). As such, it emphasizes the importance of integrating computer science, statistical science, specialist domain expertise and social science. Conway 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 (Conway, 2010). Arguably, this perspective is very similar to the interdisciplinary approach that underpins One Health and Ecohealth.

Gartner Inc, an international information technology research and advisory company, annually evaluates the maturity of emerging technologies and presents their conclusions

191 using the '*Gartner Hype Cycle*'. By representing time on the x-axis and expectations on the
192 y-axis, they define five phases through which a technology will typically pass before it
193 potentially achieves widespread adoption; starting with the Innovation Trigger phase and
194 rapidly climbing the Peak of Inflated Expectations, the cycle then descends into the Trough
195 of Disillusionment (with respect to expectations). From there it may ascend the Slope of
196 Enlightenment before finally reaching the Plateau of Productivity. As of 2014, the *Gartner*
197 *Hype Cycle* considered data science (entering the Peak of Inflated Expectations) to be lagging
198 behind both the Internet of Things (midway through the Peak) and Big Data (entering the
199 Trough of Disillusionment) (Anonymous, 2014a) - a trend that mirrors the development
200 spatial analytical methods suitable for taking advantage of the opportunities offered by
201 georeferenced Big Data.

202 ## 4. Spatial and Spatio-temporal Analysis

203 The analysis framework based on Pfeiffer et al (2008), presented in a slightly updated format
204 in Fig. 1, is still relevant for structuring the different spatial and spatio-temporal
205 epidemiological analytical methods. These are based primarily on classical statistical theory,
206 with the addition of Bayesian methods to address the issue of spatial and temporal
207 dependence. However, analysis of Big Data requires analytical algorithms which are
208 statistically robust (i.e. non-parametric) and capable of efficiently analysing very large
209 datasets. The developments for epidemiological analyses have, so far, been primarily through
210 the inclusion of machine learning regression methods as part of the modelling methods,
211 whereas in visualization and exploration it has been primarily through more effective use of
212 interfaces and flexible software environments. Below, we discuss developments for each of
213 the three analysis categories of the framework.

214 4.1 Visualization

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

221 While geographical information system (GIS) software remains at the forefront for
222 manipulating and producing complex visualisations of spatio-temporal data, the advent of
223 interactive digital maps and virtual globes such as Google Maps and Google Earth has
224 encouraged simple visualisation of disease data in real time, as illustrated by the integration
225 of such digital platforms into an ever-expanding number of animal and public-health projects
226 and platforms. For example, HealthMap (http://www.healthmap.org), together with its mobile
227 app *Outbreaks Near Me*, provides real-time surveillance of emerging public health threats
228 (Brownstein et al., 2008; Freifeld et al., 2008) while *Nature*'s use of the platform to track the
229 global spatio-temporal spread of highly pathogenic avian influenza H5N1 (Paul and White,
230 1973; Butler, 2006) won the Association of Online Publishers (AOP) Use of a New Digital
231 Platform Award in 2006.

232 Google Earth has also proved valuable for visualising disease data from informal settlements
233 or rural areas in developing countries where the lack of geolocation infrastructure such as
234 road names or house numbers precludes the use of conventional mapping software for
235 visualising disease data; in a modern day reprise of John Snow's 1856 cholera investigation,
236 use of the digital platform allowed Baker *et al*. (2011) to map the spread of a typhoid
237 outbreak in Kathmandu – where street names are not used - and trace the cause of the
238 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), 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 when, faced with only a few rudimentary topographical maps of Guinea, but no useful maps upon which to base control and surveillance efforts, 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 hours 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 Geo-Wiki a global network of volunteers working to improve the quality of global land-cover maps.

In a systematic review of visualization and analytics for infectious disease research, Carroll et al (2014) identified limitations of visualization 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 crowdsourced 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 a measure 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 visualize 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

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 visualizing 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 Modelling

Modelling approaches can be broadly categorised into data- and knowledge-driven methods (Pfeiffer et al., 2008; 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 regression methods. Knowledge-driven methods, on the other hand, require prior definition of the risk-factor variables and to define the relationship between individual risk factors and the outcome variable. 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.

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; 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), Franca and Cabral (2015) suggest

313 that, in general, tree-based regression methods tend to perform slightly better than other
314 spatial regression approaches. Requiring large datasets to be able to produce generalizable
315 inferences, these methods are ideally suited for analysing Big Data.

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

322 However, a common problem with disease regression modelling is that, while the outcome
323 variable may consist of fairly reliable disease presence information, for a usually unknown
324 number of space-time observations, absence of disease reporting may not reflect true absence
325 of disease or absence data may not be available (e.g. surveillance data). This is also common
326 in ecological species distribution modelling and has led to the development of different
327 sampling approaches to generate pseudo- absence data that can be used with regression
328 methods requiring both presence and absence data, as well as the development of specific
329 modelling techniques requiring presence-only data such as the ecological niche modelling
330 (ENM) methods including ecological niche factor analysis (ENFA), Genetic algorithm for
331 rule-set production (GARP) and maximum entropy (Maxent) (Hirzel et al., 2002; Dormann et
332 al., 2007; Elith and Leathwick, 2009; Hastie and Fithian, 2013). Requiring only disease
333 presence data means that ENM methods can make use of the extensive disease occurrence
334 data available in surveillance databases, and by extension, of web-based Big Data systems
335 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). 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 emphasizes 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. Malczeweski (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, Malczeweski (2010) and Hongoh et al (2011) emphasized 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. 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 technological changes associated with this development 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 so far have had limited impact, primarily in the area of machine learning modelling methods, but also 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 of 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 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.

406 References

407 Alvarado-Serrano, D.F., Knowles, L.L., 2014. Ecological niche models in phylogeographic
408 studies: applications, advances and precautions. Molecular ecology resources 14, 233-
409 248.

410 Anderson, C., 2008. The end of theory: The data deluge makes the scientific method
411 obsolete. Wired Magazine.

412 Andrienko, N., Andrienko, G., 2013. Visual analytics of movement: An overview of
413 methods, tools and procedures. Information Visualization 12, 3-24.

414 Anonymous, 2011. Toward Precision Medicine: Building a Knowledge Network for
415 Biomedical Research and a New Taxonomy of Disease. The National Academies
416 Press Washington, DC.

417 Anonymous, 2013. Frontiers in massive data analysis. National Research Council of the
418 National Academies, Washington DC, USA, 176pp.

419 Anonymous, 2014a. Gartner's 2014 Hype Cycle for Emerging Technologies Maps the
420 Journey to Digital Business. Gartner, Inc.

421 Anonymous, 2014b. The Internet of Things: making the most of the second digital revolution.
422 The Government Office for Science, London, UK, 38pp.

423 Baker, S., Holt, K.E., Clements, A.C.A., Karkey, A., Arjyal, A., Boni, M.F., Dongol, S.,
424 Hammond, N., Koirala, S., Duy, P.T., Nga, T.V.T., Campbell, J.I., Dolecek, C.,
425 Basnyat, B., Dougan, G., Farrar, J.J., 2011. Combined high-resolution genotyping and
426 geospatial analysis reveals modes of endemic urban typhoid fever transmission. Open
427 Biology 1.

428 Bell, G., Hey, T., Szalay, A., 2009. Computer science. Beyond the data deluge. Science 323,
429 1297-1298.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and regression trees. Wadsworth & Brooks Monterey, California, USA.

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 2009 H1N1 Influenza. New England Journal of Medicine 362, 1731-1735.

Brownstein, J.S., Freifeld, C.C., Madoff, L.C., 2009. Digital Disease Detection — Harnessing the Web for Public Health Surveillance. New England Journal of Medicine 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.

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. Clinical Infectious Diseases 49, 1557-1564.

Carrel, M., Emch, M., 2013. Genetics: A New Landscape for Medical Geography. Annals of the Association of American Geographers. Association of American Geographers 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. Journal of Biomedical Informatics 51, 287-298.

454 Chan, L.M., Brown, J.L., Yoder, A.D., 2011. Integrating statistical genetic and geospatial
455 methods brings new power to phylogeography. Molecular Phylogenetics and
456 Evolution 59, 523-537.
457 Chew, C., Eysenbach, G., 2010. Pandemics in the Age of Twitter: Content Analysis of
458 Tweets during the 2009 H1N1 Outbreak. PLoS ONE 5, e14118.
459 Chunara, R., Smolinski, M., Brownstein, J., 2013. Why We Need Crowdsourced Data in
460 Infectious Disease Surveillance. Curr Infect Dis Rep 15, 316-319.
461 Clements, A.C.A., Pfeiffer, D.U., Martin, V., 2006. Application of knowledge-driven spatial
462 modelling approaches and uncertainty management to a study of Rift Valley fever in
463 Africa. International Journal of Health Geographics 5, 57.
464 Collier, N., Doan, S., Kawazoe, A., Goodwin, R.M., Conway, M., Tateno, Y., Ngo, Q.-H.,
465 Dien, D., Kawtrakul, A., Takeuchi, K., Shigematsu, M., Taniguchi, K., 2008.
466 BioCaster: detecting public health rumors with a Web-based text mining system.
467 Bioinformatics 24, 2940-2941.
468 Collier, N., Kawazoe, A., Jin, L., Shigematsu, M., Dien, D., Barrero, R., Takeuchi, K.,
469 Kawtrakul, A., 2006. A multilingual ontology for infectious disease surveillance:
470 rationale, design and challenges. Lang Resources & Evaluation 40, 405-413.
471 Collins, F.S., Varmus, H., 2015. A New Initiative on Precision Medicine. New England
472 Journal of Medicine.
473 Conway, D., 2010. The Data Science Venn Diagram.
474 Correa, T.R., Assuncao, R.M., Costa, M.A., 2014. A critical look at prospective surveillance
475 using a scan statistic. Stat Med.
476 Costa, M., Kulldorff, M., 2014. Maximum linkage space-time permutation scan statistics for
477 disease outbreak detection. International Journal of Health Geographics 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. Journal of the American Veterinary Medical Association 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.

Dhar, V., 2013. Data science and prediction. Commun. ACM 56, 64-73.

Dodge, M., Kitchin, R., 2013. Crowdsourced cartography: mapping experience and knowledge. Environment and Planning A 45, 19-36.

Dormann, C.F., McPherson, J.M., Araújo, M.B., Bivand, R., Bolliger, J., Carl, G., Davies, R.G., Hirzel, A., Jetz, W., Daniel Kissling, W., Kühn, I., Ohlemüller, R., Peres-Neto, P.R., Reineking, B., Schröder, B., Schurr, F.M., Wilson, R., 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography 30, 609-628.

Eastman, J.R., 2009. Decision Support: Uncertainty Management. IDRISI Guide to GIS and Image Processing. Accessed in IDRISI Andes., Worcester, MA: Clark University, 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.

Elith, J., Leathwick, J.R., 2009. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. Annual Review of Ecology, Evolution, and Systematics 40, 677-697.

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.

503 Faria, N.R., Suchard, M.A., Rambaut, A., Lemey, P., 2011. Toward a quantitative
504 understanding of viral phylogeography. Current opinion in virology 1, 423-429.
505 Feizizadeh, B., Jankowski, P., Blaschke, T., 2014. A GIS based spatially-explicit sensitivity
506 and uncertainty analysis approach for multi-criteria decision analysis. Computers &
507 Geosciences 64, 81-95.
508 Fernández, A., del Río, S., López, V., Bawakid, A., del Jesus, M.J., Benítez, J.M., Herrera,
509 F., 2014. Big Data with Cloud Computing: an insight on the computing environment,
510 MapReduce, and programming frameworks. Wiley Interdisciplinary Reviews: Data
511 Mining and Knowledge Discovery 4, 380-409.
512 Firestone, S.M., Christley, R.M., Ward, M.P., Dhand, N.K., 2012. Adding the spatial
513 dimension to the social network analysis of an epidemic: investigation of the 2007
514 outbreak of equine influenza in Australia. Prev Vet Med 106, 123-135.
515 Firestone, S.M., Ward, M.P., Christley, R.M., Dhand, N.K., 2011. The importance of location
516 in contact networks: Describing early epidemic spread using spatial social network
517 analysis. Prev Vet Med 102, 185-195.
518 França, S., Cabral, H.N., 2015. Predicting fish species richness in estuaries: Which modelling
519 technique to use? Environmental Modelling & Software 66, 17-26.
520 Freifeld, C.C., Mandl, K.D., Reis, B.Y., Brownstein, J.S., 2008. HealthMap: global infectious
521 disease monitoring through automated classification and visualization of Internet
522 media reports. J. Am. Med. Inform. Assoc. 15, 150-157.
523 Gandomi, A., Haider, M., 2015. Beyond the hype: Big data concepts, methods, and analytics.
524 International Journal of Information Management 35, 137-144.
525 Giebultowicz, S., Ali, M., Yunus, M., Emch, M., 2011. The simultaneous effects of spatial
526 and social Networks on cholera transmission. Interdisciplinary Perspectives on
527 Infectious Diseases 2011.

528 Gilbert, M., Golding, N., Zhou, H., Wint, G.R., Robinson, T.P., Tatem, A.J., Lai, S., Zhou,
529 S., Jiang, H., Guo, D., Huang, Z., Messina, J.P., Xiao, X., Linard, C., Van Boeckel,
530 T.P., Martin, V., Bhatt, S., Gething, P.W., Farrar, J.J., Hay, S.I., Yu, H., 2014.
531 Predicting the risk of avian influenza A H7N9 infection in live-poultry markets across
532 Asia. Nature communications 5, 4116.
533 Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L., 2009a.
534 Detecting influenza epidemics using search engine query data. Nature 457, 1012-
535 1014.
536 Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L., 2009b.
537 Detecting influenza epidemics using search engine query data. Nature 457, 1012-
538 1014.
539 Goodchild, M., 2007. Citizens as sensors: the world of volunteered geography. GeoJournal
540 69, 211-221.
541 Goodchild, M.F., Li, L., 2012. Assuring the quality of volunteered geographic information.
542 Spatial Statistics 1, 110-120.
543 Grein, T.W., Kamara, K.B., Rodier, G., Plant, A.J., Bovier, P., Ryan, M.J., Ohyama, T.,
544 Heymann, D.L., 2000. Rumors of disease in the global village: outbreak verification.
545 Emerg Infect Dis 6, 97–102.
546 Haklay, M., 2010. How good is volunteered geographical information? A comparative study
547 of OpenStreetMap and Ordnance Survey datasets. Environment and planning. B,
548 Planning & design 37, 682.
549 Hartley, D.M., Nelson, N.P., Arthur, R.R., Barboza, P., Collier, N., Lightfoot, N., Linge, J.P.,
550 van der Goot, E., Mawudeku, A., Madoff, L.C., Vaillant, L., Walters, R., Yangarber,
551 R., Mantero, J., Corley, C.D., Brownstein, J.S., 2013. An overview of Internet
552 biosurveillance. Clinical Microbiology and Infection 19, 1006-1013.

553 Hastie, T., Fithian, W., 2013. Inference from presence-only data; the ongoing controversy.
554 Ecography 36, 864-867.
555 Hay, S.I., George, D.B., Moyes, C.L., Brownstein, J.S., 2013. Big data opportunities for
556 global infectious disease surveillance. PLoS Med 10, e1001413.
557 Hay, S.I., Graham, A., Rogers, D.J., 2006. Global mapping of infectious diseases: Methods,
558 examples and emerging applications. Academic Press Amsterdam, The Netherlands.
559 Heipke, C., 2010. Crowdsourcing geospatial data. ISPRS Journal of Photogrammetry and
560 Remote Sensing 65, 550-557.
561 Heymann, D.L., Rodier, G.R., 2001. Hot spots in a wired world: WHO surveillance of
562 emerging and re-emerging infectious diseases. The Lancet Infectious Diseases 1, 345-
563 353.
564 Hirzel, A.H., Hausser, J., Chessel, D., Perrin, N., 2002. Ecological-niche factor analysis: How
565 to compute habitat-suitability maps without absence data? Ecology 83, 2027-2036.
566 Hirzel, A.H., Le Lay, G., Helfer, V., Randin, C., Guisan, A., 2006. Evaluating the ability of
567 habitat suitability models to predict species presences. Ecological Modelling 199,
568 142-152.
569 Hodson, H., 2014. Online army helps map Guinea's Ebola outbreak. New Scientist.
570 Hongoh, V., Hoen, A.G., Aenishaenslin, C., Waaub, J.P., Belanger, D., Michel, P., Lyme,
571 M.C., 2011. Spatially explicit multi-criteria decision analysis for managing vector-
572 borne diseases. Int J Health Geogr 10, 70.
573 http://www.google.org/flutrends, G.F.T.
574 Jankowski, P., Fraley, G., Pebesma, E., 2014. An exploratory approach to spatial decision
575 support. Computers, Environment and Urban Systems 45, 101-113.
576 Kambatla, K., Kollias, G., Kumar, V., Grama, A., 2014. Trends in big data analytics. Journal
577 of Parallel and Distributed Computing 74, 2561-2573.

578 Kamel Boulos, M., Al-Shorbaji, N., 2014. On the Internet of Things, smart cities and the
579 WHO Healthy Cities. International Journal of Health Geographics 13, 10.
580 Kamel Boulos, M., Resch, B., Crowley, D., Breslin, J., Sohn, G., Burtner, R., Pike, W.,
581 Jezierski, E., Chuang, K.-Y., 2011. Crowdsourcing, citizen sensing and sensor web
582 technologies for public and environmental health surveillance and crisis management:
583 trends, OGC standards and application examples. International Journal of Health
584 Geographics 10, 67.
585 Keller, M., Blench, M., Tolentino, H., Freifeld, C., Mandl, K., Mawudeku, A., Eysenbach,
586 G., Brownstein, J., 2009. Use of unstructured event-based reports for global infectious
587 disease surveillance. Emerg Infect Dis 15, 689-695.
588 Lawson, A.B., 2014. Hierarchical modeling in spatial epidemiology. Wiley Interdisciplinary
589 Reviews: Computational Statistics 6, 405-417.
590 Lazer, D., Kennedy, R., King, G., Vespignani, A., 2014. Big data. The parable of Google Flu:
591 traps in big data analysis. Science 343, 1203-1205.
592 Lee, J., Kao, H.-A., Yang, S., 2014. Service Innovation and Smart Analytics for Industry 4.0
593 and Big Data Environment. Procedia CIRP 16, 3-8.
594 Leetaru, K., 2014. Why big data missed the early warning signs of ebola. Foreign Policy.
595 Li, J., Heap, A.D., 2014. Spatial interpolation methods applied in the environmental sciences:
596 A review. Environmental Modelling & Software 53, 173-189.
597 Liang, L., Xu, B., Chen, Y., Liu, Y., Cao, W., Fang, L., Feng, L., Goodchild, M.F., Gong, P.,
598 2010. Combining Spatial-Temporal and Phylogenetic Analysis Approaches for
599 Improved Understanding on Global H5N1 Transmission. PloS one 5.
600 Ligmann-Zielinska, A., Jankowski, P., 2014. Spatially-explicit integrated uncertainty and
601 sensitivity analysis of criteria weights in multicriteria land suitability evaluation.
602 Environmental Modelling & Software 57, 235-247.

603 Malak, M., 2014. The Fourth Bubble in the Data Science Venn Diagram: Social Sciences.
604 Malczewski, J., 2006. GIS-based multicriteria decision analysis: a survey of the literature.
605 International Journal of Geographical Information Science 20.
606 Malczewski, J., 2010. Multiple criteria decision analysis and geographic information systems.
607 In: Ehrgott, M., Greco, S., Figueira, J.R. (Eds.), Trends in multiple criteria decision
608 analysis. Springer, New York, 369-395.
609 Martin, V., Pfeiffer, D.U., Zhou, X., Xiao, X., Prosser, D.J., Guo, F., Gilbert, M., 2011.
610 Spatial Distribution and Risk Factors of Highly Pathogenic Avian Influenza (HPAI)
611 H5N1 in China. PLoS Pathog 7, e1001308.
612 Mayer-Schönberger, V., Cukier, K., 2014. Big data : a revolution that will transform how we
613 live, work, and think. Mariner Books, Houghton Mifflin Harcourt Boston.
614 Michel, J.B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., Google Books, T., Pickett, J.P.,
615 Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M.A., Aiden,
616 E.L., 2011. Quantitative analysis of culture using millions of digitized books. Science
617 331, 176-182.
618 Milinovich, G.J., Magalhães, R.J.S., Hu, W., 2015. Role of big data in the early detection of
619 Ebola and other emerging infectious diseases. The Lancet Global Health 3, e20-e21.
620 Milinovich, G.J., Williams, G.M., Clements, A.C.A., Hu, W., 2014. Internet-based
621 surveillance systems for monitoring emerging infectious diseases. The Lancet
622 Infectious Diseases 14, 160-168.
623 Mortari, A., Lorenzelli, L., 2014. Recent sensing technologies for pathogen detection in milk:
624 a review. Biosensors & bioelectronics 60, 8-21.
625 Murray, A.T., Grubesic, T.H., Wei, R., 2014. Spatially significant cluster detection. Spatial
626 Statistics 10, 103-116.

627 Mykhalovskiy, E., Weir, L., 2006. The Global Public Health Intelligence Network and early
628 warning outbreak detection: a Canadian contribution to global public health. Can J
629 Public Health 97, 42-44.
630 Norvig, P., 2009. All we want are the facts, ma'am.
631 O'Driscoll, A., Daugelaite, J., Sleator, R.D., 2013. 'Big data', Hadoop and cloud computing
632 in genomics. Journal of Biomedical Informatics 46, 774-781.
633 Okabe, A., Sugihara, K., 2012. Spatial analysis along networks - Statistical and
634 computational methods. John Wiley & Sons Chichester, UK.
635 Oliver, M.A., Webster, R., 2014. A tutorial guide to geostatistics: Computing and modelling
636 variograms and kriging. CATENA 113, 56-69.
637 Osborne, P.E., Alonso, J.C., Bryant, R.G., 2001. Modelling landscape-scale habitat use using
638 GIS and remote sensing: a case study with great bustards. Journal of Applied Ecology
639 38, 458-471.
640 Osei-Bryson, K.M., 2003. Supporting knowledge elicitation and consensus building for
641 Dempster-Shafer decision models. International Journal of Intelligent Systems 18,
642 129-148.
643 Ostle, A.G., Murtle, D.C., Welter, C.J., 1986. Reducing the effects of enzootic pneumonia
644 and atrophic rhinitis. Veterinary Medicine, 772-775.
645 Paul, J.R., White, C., 1973. Serological Epidemiology. 19-55.
646 Peters, D.P.C., Havstad, K.M., Cushing, J., Tweedie, C., Fuentes, O., Villanueva-Rosales, N.,
647 2014. Harnessing the power of big data: infusing the scientific method with machine
648 learning to transform ecology. Ecosphere 5, art67.
649 Pfeiffer, D.U., Robinson, T.P., Stevenson, M., Stevens, K.B., Rogers, D.J., Clements, A.C.A.,
650 2008. Spatial analysis in epidemiology. Oxford University Press Oxford, UK.
651 Pigliucci, M., 2009. The end of theory in science? EMBO reports 10, 534.

652 Pigott, D.M., Golding, N., Mylne, A., Huang, Z., Henry, A.J., Weiss, D.J., Brady, O.J.,
653 Kraemer, M.U., Smith, D.L., Moyes, C.L., Bhatt, S., Gething, P.W., Horby, P.W.,
654 Bogoch, II, Brownstein, J.S., Mekaru, S.R., Tatem, A.J., Khan, K., Hay, S.I., 2014.
655 Mapping the zoonotic niche of Ebola virus disease in Africa. eLife 3, e04395.
656 Porter, J.H., Hanson, P.C., Lin, C.C., 2012. Staying afloat in the sensor data deluge. Trends
657 Ecol Evol 27, 121-129.
658 Pouilly, F., Viel, J.F., Mialot, J.P., Sanaa, M., Humblot, P., Ducrot, C., Grimard, B., 1994.
659 Risk Factors for post-partum Anoestrus in Charolais Beef Cows in France. Preventive
660 Veterinary Medicine 18, 305-314.
661 Prates, M.O., Kulldorff, M., Assuncao, R.M., 2014. Relative risk estimates from spatial and
662 space-time scan statistics: are they biased? Stat Med 33, 2634-2644.
663 Pybus, O.G., Suchard, M.A., Lemey, P., Bernardin, F.J., Rambaut, A., Crawford, F.W., Gray,
664 R.R., Arinaminpathy, N., Stramer, S.L., Busch, M.P., Delwart, E.L., 2012. Unifying
665 the spatial epidemiology and molecular evolution of emerging epidemics. Proc Natl
666 Acad Sci U S A 109, 15066-15071.
667 Rutten, C.J., Velthuis, A.G., Steeneveld, W., Hogeveen, H., 2013. Invited review: sensors to
668 support health management on dairy farms. J Dairy Sci 96, 1928-1952.
669 Sanchez-Matamoros, A., Sanchez-Vizcaino, J.M., Rodriguez-Prieto, V., Iglesias, E.,
670 Martinez-Lopez, B., 2014. Identification of Suitable Areas for African Horse Sickness
671 Virus Infections in Spanish Equine Populations. Transbound Emerg Dis.
672 Sarojinie Fernando, W.T.P., Hazelton, M.L., 2014. Generalizing the spatial relative risk
673 function. Spatial and Spatio-temporal Epidemiology 8, 1-10.
674 Schutt, R., O'Neil, C., 2013. Doing data science. O'Reilly Media Sebastopol, California,
675 USA.

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.

St Louis, C., Zorlu, G., 2012. Can Twitter predict disease outbreaks?

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. Spatial and spatio-temporal epidemiology 4, 1-14.

Stevens, K.B., Pfeiffer, D.U., 2011. Spatial modelling of disease using data- and knowledge-driven approaches. Spat.Spatiotemporal.Epidemiol. 2, 125-133.

Tatem, A.J., 2014. Mapping population and pathogen movements. International 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. Malaria journal 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. Advances in Disease 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

700 evaluation to map areas at risk for Rift Valley fever vector-borne transmission in
701 Italy. Transbound Emerg Dis 60 Suppl 2, 14-23.
702 van Zyl, T., 2014a. Algorithmic considerations for geospatial and/or temporal big data. In:
703 Karimi, H.A. (Ed.), Big Data - Techniques and technologies in geoinformatics. CRC
704 Press, Boca Raton, Florida, USA, 117-132.
705 van Zyl, T., 2014b. Machine learning on geospatial big data. In: Karimi, H.A. (Ed.), Big Data
706 - Techniques and technologies in geoinformatics. CRC Press, Boca Raton, Florida,
707 USA, 133-148.
708 Vatsavai, R.R., Ganguly, A., Chandola, V., Stefanidis, A., Klasky, S., Shekhar, S., 2012.
709 Spatiotemporal data mining in the era of big spatial data: algorithms and applications.
710 Proceedings of the 1st ACM SIGSPATIAL International Workshop on Analytics for
711 Big Geospatial Data. ACM, Redondo Beach, California, 1-10.
712 Wesolowski, A., Buckee, C.O., Bengtsson, L., Wetter, E., Lu, X., Tatem, A.J., 2014a.
713 Commentary: Containing the Ebola outbreak ? the potential and challenge of mobile
714 network data. PLoS Currents Outbreaks Edition 1.
715 Wesolowski, A., Stresman, G., Eagle, N., Stevenson, J., Owaga, C., Marube, E., Bousema,
716 T., Drakeley, C., Cox, J., Buckee, C.O., 2014b. Quantifying travel behavior for
717 infectious disease research: a comparison of data from surveys and mobile phones.
718 Scientific reports 4, 5678.
719 Wilson, K., Brownstein, J.S., 2009. Early detection of disease outbreaks using the Internet.
720 Canadian Medical Association Journal 180, 829-831.
721 Wilson, K., von Tigerstrom, B., McDougall, C., 2008. Protecting global health security
722 through the International Health Regulations: requirements and challenges. Canadian
723 Medical Association Journal 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. Euro Surveill 14, pii=19386.

Wing, J.M., 2008. Computational thinking and thinking about computing. Philosophical transactions. Series A, Mathematical, physical, and engineering sciences 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.

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., 2014. Mining data with random forests: current options for real-world applications. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 4, 55-63.