Spatial and Spatio-temporal Epidemiology 13 (2015) 15–29

Contents lists available at ScienceDirect

Spatial and Spatio-temporal Epidemiology

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

Review Article

Sources of spatial animal and human health data: Casting the net wide to deal more effectively with increasingly complex disease problems

Kim B. Stevens *, Dirk U. Pfeiffer

Veterinary Epidemiology, Economics and Public Health Group, Dept. of Production & Population Health, Royal Veterinary College, London, United Kingdom

ARTICLE INFO

Article history:
Received 2 March 2015
Accepted 28 April 2015
Available online 8 May 2015

Keywords:
Big data
Data warehouse
Google Earth
mHealth
Spatial data
Volunteered geographic information

ABSTRACT

During the last 30 years it has become commonplace for epidemiological studies to collect locational attributes of disease data. Although this advancement was driven largely by the introduction of handheld global positioning systems (GPS), and more recently, smartphones and tablets with built-in GPS, the collection of georeferenced disease data has moved beyond the use of handheld GPS devices and there now exist numerous sources of crowdsourced georeferenced disease data such as that available from georeferencing of Google search queries or Twitter messages. In addition, cartography has moved beyond the realm of professionals to crowdsourced mapping projects that play a crucial role in disease control and surveillance of outbreaks such as the 2014 West Africa Ebola epidemic. This paper provides a comprehensive review of a range of innovative sources of spatial animal and human health data including data warehouses, mHealth, Google Earth, volunteered geographic information and mining of internet-based big data sources such as Google and Twitter. We discuss the advantages, limitations and applications of each, and highlight studies where they have been used effectively.

1. Introduction

Over the last 30 years it has become commonplace for epidemiological studies or surveys to collect locational (spatial) attributes for disease data (Pfeiffer et al., 2008). Although this advancement has been driven largely by the introduction of handheld global positioning systems (GPS), and more recently, smartphones and tablet computers with built-in GPS that facilitate geo-tagged data collection, it also highlights the increased awareness of the importance of the spatial aspect when developing efficacious animal disease surveillance and control strategies (Table 1). Unfortunately, as a result of the particular challenges currently facing health workers and researchers, for spatial disease data to be able to effectively inform innovative surveillance and disease control strategies, it needs to move beyond the fundamentals of collecting geo-referenced disease event data in individual studies and instead focus on an inclusive approach that Eysenbach (2001), in his definition of eHealth, referred to as '*a state-of-mind, a way of thinking, an attitude, and a commitment for networked, global thinking, to improve health care locally, regionally, and worldwide by using information and communication technology*'.

This collective, crowdsourced approach was aptly illustrated 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, Médecins Sans Frontières (MSF) personnel

* Corresponding author at: Royal Veterinary College, Hawkshead Lane, North Mymms, Hatfield, Hertfordshire AL9 7TA, United Kingdom.

E-mail addresses: kstevens@rvc.ac.uk (K.B. Stevens), pfeiffer@rvc.ac.uk (D.U. Pfeiffer).

http://dx.doi.org/10.1016/j.sste.2015.04.003
1877-5845/

Table 1
Using spatial analysis to inform risk-based animal disease surveillance and control.

Mapping disease distribution	Disease distribution maps range from simple dot maps showing the location of disease events to predictive risk maps created using statistical algorithms that combine disease occurrence data with environmental covariates (Pigott et al., 2014). But no matter what form they take, visualizing the spatial pattern of disease – be it at a global, national or local scale – is fundamental for informing risk-based disease surveillance and control strategies in several ways. Simple visualizations allow the extent of the disease to be delineated and disease frequency monitored, and when combined with maps of environmental factors or those highlighting the spatially heterogeneous distribution of at-risk populations, they can also be used to estimate disease burden (Hay et al., 2010; Robinson et al., 2002) and identify target populations for intervention (Tatem et al., 2011; Guerra et al., 2010, 2008, 2006). Visualizing disease distribution can also be fundamental in directing control and elimination efforts. Clements et al. (2013) describe how measures to eliminate malaria from endemic countries have generally adopted a spatially progressive elimination approach referred to as *shrinking the malaria map* in which eradication efforts initially focus on the geographical perimeter of endemic areas and work inwards, effectively localizing disease distribution which allows for more efficient treatment and control (Feachem et al., 2010) Apart from the key role maps play in informing risk-led decision making, they also serve a more practical purpose such as facilitating integration and synthesis of data from a wide range of diverse sources, each possibly capturing information about disease and relevant risk factors at different scales (Bergquist and Tanner, 2012; Bennema et al., 2014). As a result, cartographers need to decide on the most appropriate scale at which to present the data for it to be useful; data presented at administrative level 1 (province or region) inevitably cannot capture the fine-scale heterogeneity of most infection patterns and so estimates of numbers of individuals requiring treatment tend to be incorrect (Brooker et al., 2010).
Cluster detection	A clustered spatial arrangement of disease events suggests the presence of a contagious process or localised risk factor. Apart from the fact that spatial targeting of interventions at high-risk areas is more cost-effective than uniform resource allocation (Stark et al., 2006) and therefore such identification is essential for informing risk-based disease surveillance and control efforts. Identification of significant disease clusters can also advance our understanding of a disease in several ways including suggesting potential risk factors for further investigation either directly (Calistri et al., 2013; French et al., 2005; Sinkala et al., 2014; Kelen et al., 2012; Nogareda et al., 2013; Poljak et al., 2007; Le et al., 2012; Vigre et al., 2005; Ward and Carpenter, 2000), or indirectly when analysis of model residuals indicates the modelled predictors do not explain fully the spatial heterogeneity in disease distribution (Méroc et al., 2014; Borba et al., 2013), or by defining the scale of disease clustering (French et al., 2005; Le et al., 2012; French et al., 1999; Wilesmith et al., 2003; Picado et al., 2007; Picado et al., 2011; Porphyre et al., 2007; Sanchez et al., 2005; Minh et al., 2009; Minh et al., 2010; Xu et al., 2012; Métras et al., 2012; Abatih and Ersbøll, 2009) and thereby indicate likely transmission mechanisms involved in disease spread (Sinkala et al., 2014; Ward et al., 2013; Loobuyck et al., 2009; Ohlson et al., 2014; Rosendal et al., 2014; Poljak et al., 2010). Cluster detection can also be used identify areas where vectors and hosts coincide resulting in potentially increased risk of disease transmission (Shaman, 2007; Hennebelle et al., 2013; Swirski et al., 2007), highlight possible regional differences in disease transmission (Kelen et al., 2012), or track the direction and geographical extent of disease spread (Wilesmith et al., 2003; Denzin et al., 2013; Lian et al., 2007)
Spatial modelling	Spatial modelling techniques can be divided into data- and knowledge-driven methods (Stevens and Pfeiffer, 2011), the former characterised by the use of statistical methods for defining relationships between risk factors and disease risk, while knowledge-driven modelling approaches are based on existing knowledge about the causal relationships associated with the disease risk of interest. Statistical analysis is used to generate data-driven models from information collected through surveillance and other means. Such models generate quantitative estimates of risk and the relative weights of risk factors. The results of such models are used for a variety of purposes including targeting areas for disease surveillance, risk management, simulating different control scenarios, or predicting what will happen under different environmental conditions such as those resulting from climate change (i.e. temporal prediction), or identifying new geographical areas suitable for the introduction of diseases (i.e. spatial prediction)

enlisted the help of the Humanitarian OpenStreetMap Team (HOT) 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.

In addition to this collective approach, for spatial disease data to be effective in the 21st century, it needs to meet certain requirements. Firstly, the increasing number of transboundary disease epidemics has emphasized the need for animal and human health information systems that are no longer circumscribed by regional or national borders; transparent collection and sharing of disease data needs to occur at a global scale. Secondly globalization has substantially increased the speed and magnitude of disease spread. In the 2001 UK foot and mouth disease (FMD) outbreak it was estimated that at least 57 premises from 16 counties were infected before the first case was reported (Gibbens and Wilesmith, 2002) while in 2007, equine influenza spread rapidly throughout two Australian states as a result of infected horses attending an equestrian event (Cowled et al., 2009); approximately 70,000 horses on over 9000 premises were infected with most of the geographic dissemination occurring within the first ten days of the epidemic. For containment to be effective, reporting of disease events needs to be as rapid as possible. This is of particular concern in developing countries where reporting of animal disease events can be delayed by months (Karimuribo et al., 2012) while lag times for such reports as the Centers for Disease Control and Prevention (CDC) US Influenza Sentinel Provider Surveillance reports are currently in the order of 1–2 weeks (Ginsberg et al., 2009).

During the past decade, collecting spatial disease data has moved beyond the use of handheld GPS devices and there now exist numerous sources of crowdsourced georeferenced disease data such as that available from georeferencing Google search queries or Twitter messages. Not surprisingly, the focus so far has been on human health,