Preventive Veterinary Medicine 120 (2015) 131–140

Contents lists available at ScienceDirect

Preventive Veterinary Medicine

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

An approach for de-identification of point locations of livestock premises for further use in disease spread modeling

Michael K. Martin[a,*], Julie Helm[a], Kelly A. Patyk[b]

[a] Livestock Poultry Health Division, Clemson University, Columbia, SC 29224, USA
[b] U.S Department of Agriculture, Animal and Plant Health Inspection Service, Veterinary Services, Science Technology and Analysis Services, Center for Epidemiology and Animal Health, 2150 Centre Avenue, Building B, Fort Collins, CO 80526, USA

ARTICLE INFO

Article history:
Received 21 April 2014
Received in revised form 31 March 2015
Accepted 17 April 2015

Keywords:
Modeling
Simulation
K-anonymity
Privacy
Confidentiality
Random permutation

ABSTRACT

We describe a method for de-identifying point location data used for disease spread modeling to allow data custodians to share data with modeling experts without disclosing individual farm identities. The approach is implemented in an open-source software program that is described and evaluated here. The program allows a data custodian to select a level of de-identification based on the K-anonymity statistic. The program converts a file of true farm locations and attributes into a file appropriate for use in disease spread modeling with the locations randomly modified to prevent re-identification based on location. Important epidemiological relationships such as clustering are preserved to as much as possible to allow modeling similar to those using true identifiable data. The software implementation was verified by visual inspection and basic descriptive spatial analysis of the output. Performance is sufficient to allow de-identification of even large data sets on desktop computers available to any data custodian.

1. Introduction

Spatially explicit, stochastic disease spread models such as the North American Animal Disease Spread Model (NAADSM) (Harvey et al., 2007), AusSpread (Garner and Beckett, 2005; Beckett and Garner, 2007) and InterSpread Plus (Stevenson et al., 2013) are powerful tools that can be used to assist in the decision-making process for animal health and disease management policy. Simulation models have been used to assess disease behavior under a variety of conditions, to compare the efficacy of control measures, to develop and evaluate contingency plans, and to estimate resources required in the event of an outbreak (Ward et al., 2009a, 2009b; Tildesley et al., 2011; Porphyre et al., 2013; Roche et al., 2014). Disease spread models apply input data through the use of stochastic processes to simulate the spread of disease from infected livestock premises to susceptible ones. The probability of spread depends on a number of model parameters along with attributes of the premises involved. Premises may be classified by type and number of livestock present. If a premises produces livestock such as poultry or swine for a vertically integrated production company, the specific company affiliation is an important attribute for modeling due to connectivity between vertically integrated premises. A key factor in the probability of spread is the distance – and in some cases direction – between premises. Therefore, the spatial location of each premises is important to model outputs and the conclusions that can be drawn from them. These models take as one of their input sets the true point or area locations of the premises in the populations at risk. Confident use of model outputs to inform decision-making is dependent upon the quality of model inputs, including animal population data. The spatial scales and distributions of animal populations used as model inputs to simulate disease transmission and control can influence model outputs such as those describing disease dynamics, outbreak severity and extent (Highfield et al., 2008; Tildesley et al., 2010; Carpenter, 2011; Reeves, 2012). The true location of each livestock premises is the ideal data source. Unfortunately, such data are not always available to modelers due to confidentiality requirements and privacy concerns. In the United States, for example, very few data exist regarding the locations of livestock premises, and those data that do exist are often not accessible. Various government agencies and private industry groups have real-world premises location and population data. In virtually every case such data were collected and are held under formal or informal promises of confidentiality (USDA, 1985–2002; USDA–NASS, 2007; USDA–APHIS, 2012, Title 7 Section 2276). Custodians of these data are under dual obligations to use these data constructively for the prevention and control

[*] Corresponding author at: Livestock Poultry Health Division, Clemson University, PO Box 102406, Columbia, SC 29224-2406, USA. Tel.: +1 803 788 2260; fax: +1 803 788 8058.
E-mail address: mmarti5@clemson.edu (M.K. Martin).

http://dx.doi.org/10.1016/j.prevetmed.2015.04.010

of animal disease while restricting their unauthorized release for other purposes. In many cases, the expertise needed to use the data for disease spread modeling lies outside the data custodian organization. In these cases there is a need to disclose the location data in a way that allows disease spread modeling while maintaining an adequate assurance of privacy for the individual premises represented. The words privacy and confidentiality are used here in closely related but slightly different ways. Privacy is the right of individuals to control disclosure of information about themselves. Confidentiality is the obligation of second party data custodians to control further disclosure of information about individuals who have legitimate privacy concerns.

1.1. Sources of farm premises data commonly used in modeling

Because true location data are so seldom available to disease spread modelers, particularly in the United States, geographic coordinates and accompanying premises attributes are often simulated from aggregate data. The most common source of surrogate premises location data for the United States is the U.S. Department of Agriculture's Census of Agriculture (USDA, 1994). The National Agricultural Statistics Service (NASS) collects a detailed census of farms every 5 years. However, NASS is prohibited from disclosing information that allows the identification of the person that supplied the information. To protect privacy these data are made available as county-level totals for numbers of farms and numbers of animals for different types of livestock and poultry.

Various techniques have been used to simulate premises locations and animal holdings from county-level summary data (Melius et al., 2006; Melius, 2007; Bruhn et al., 2012; Tildesley and Ryan, 2012). Several studies have evaluated variations in model outcomes when different sources of population data are used. Model outcomes are sensitive to spatial clustering, distribution of farm populations such as some areas having a few large farms and many small ones versus other areas having mostly medium sized farms, and type and species composition of each farm (Reeves, 2012; Tildesley and Ryan, 2012). Model parameters can be adjusted to match real-world outbreak data in order to compensate for the lack of realistic clustering in synthetic data sources. It has been shown that such re-parameterized models are effective in optimizing control strategies. However, because this approach depends on real-world outbreak data, it can only be used reactively and is therefore not applicable when using models in advance of an outbreak for planning and preparedness (Tildesley et al., 2010).

1.2. De-identification

Given the importance of quality population data for use in models, one solution (as an alternative to synthesized or randomly placed farms) is to use the true locations in a way that allows modelers access to all the epidemiologically important attribute and spatial relationships between the premises while still preventing the identification of individual premises in the real world from the modeling data. Preventing identification while retaining important epidemiological information is known as "de-identification". A side benefit of this solution is that it produces data that appear realistic to the decision-makers who will be called upon to consider the modeling outcomes. While this is completely separate from the mathematical conclusions drawn from the modeling, it may build confidence in model outputs.

Since the passage of the security and privacy provisions of the Health Insurance Portability and Accountability Act there has been a great body of research on how to render medical data such that the "risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information" (45 CFR Subtitle A, Subchapter C, Part 164–514). A few basic techniques have been used to de-identify medical records. These include aggregation, generalization, and suppression (Domingo-Ferrer and Torra, 2005). An additional approach involves random perturbation of the data in a way that preserves its value for statistical analysis (Adam and Worthmann, 1989).

The NASS Census of Agriculture provides examples of aggregation and suppression. Aggregation is applied in that only county-level totals are released to the public. Even then, some counties have so few farms of certain types that there would be an unacceptable loss of privacy. So NASS uses suppression in addition to aggregation. To prevent unacceptable re-identification risk, NASS removes even the aggregate totals from the published data for those counties. They replace the data with the symbol (D) (USDA–NASS, 2007).

In other situations generalization is used to avoid completely suppressing some information, such as date of birth, by replacing it with more general – and therefore less specific to any single individual – representations of the same information such as age group or year of birth. Generalization can be used with categorical attributes such as types of animals by combining very small groups that might result in compromised privacy.

Random perturbation is a statistical de-identification method in which a certain amount of random variation is added to individual data values. This perturbation changes those values enough to reduce analysts' ability to identify the source while preserving essential statistics in the overall epidemiologic analysis of the data.

1.3. K-anonymity

Generalization, suppression, and random perturbation can be applied to individual records, also known as "microdata". The risk is that the collection of information that is released may be combined with other generally available information such as phone directories, voter rolls, etc. to uniquely identify the individual. Efforts to quantify that risk lead to the development of the K-anonymity measure (Samarati and Sweeney, 1998).

The "K" in K-anonymity refers to the minimum number of distinct individuals whose identities would be indistinguishable from each other in a released dataset by linking with generally available records. If, for example, for each premises in a dataset we have removed or generalized each fact about that premises sufficiently that no one can tell it apart from at least five other premises then we say that the record has a K-anonymity of five. The value of five for K is something of a historical artifact similar to the way 0.05 has become the default threshold of statistical significance in biology. Various values from 3 to 15 have been suggested (El Emam and Dankar, 2008).

The first step in de-identification is to remove any directly identifying information such as names and Social Security numbers. The required level of K-anonymity is then achieved by controlling the quasi-identifiers remaining in the released data. Quasi-identifiers are data elements that may be paired with other records used to attempt re-identification. These elements would include the obvious things like name and address but also any other facts that might appear in other data sources that identify the premises. In the data used for disease spread modeling the quasi-identifiers include both the information about the premises such as the types of animals and integrator company affiliation, and their spatial locations.

In addition to quasi-identifiers the data set to be de-identified may contain sensitive information or there would be no point in de-identifying it. These are known as "confidential outcome attributes" (Domingo-Ferrer and Torra, 2005). If within a K-group these sensitive data are essentially the same, someone with the set, while not knowing the identity of any one individual, knows the sensitive fact