Title

An Early Warning System for Detecting H1N1 Disease Outbreak - A Spatio-temporal Approach

Authors

Lai Poh-Chin[1], Chow Chun Bong[2], Wong Ho Ting[3], Kwong Kim Hung[4], Kwan Yat Wah[5], Liu Shao Haei[6], Tong Wah Kun[7], Cheung Wai Keung[8], Wong Wing Leung[9]

[1] Associate Professor, Department of Geography, The University of Hong Kong, Pokfulam Road, Hong Kong SAR, China. Email: pclai@hku.hk

[2] Hon. Consultant, Former Medical Director, Hospital Authority Infectious Disease Centre, Princess Margaret Hospital, Hong Kong SAR, China. Email: chowcb@netvigator.com

[3] Research Associate, CUHK Jockey Club Institute of Ageing, The Chinese University of Hong Kong, Hong Kong SAR, China. Email: frankwong@connect.hku.hk

[4] Former Postdoctoral Fellow, Department of Geography, The University of Hong Kong, Pokfulam Road, Hong Kong SAR, China. Email: h0110454@graduate.hku.hk

[5] Associate Consultant, Department of Paediatrics & Adolescent Medicine, Princess Margaret Hospital, Hong Kong SAR, China. Email: kwanyw1@ha.org.hk

[6] Chief Manager, Hospital Authority, Kowloon, Hong Kong. Email: liush@ha.org.hk

[7] Departmental Operation Manager, Hospital Authority Infectious Centre, Princess Margaret Hospital, Hong Kong SAR, China. Email: tongwk2@ha.org.hk

[8] Assistant Professor, School of Nursing, The University of Hong Kong, Pokfulam Road, Hong Kong SAR, China. Email: lawrencewkcheung@gmail.com

[9] Research Fellow, City University of Hong Kong, Hong Kong SAR, China. Email: wingleungwong@hotmail.com.hk

Abstract

The outbreaks of new and emerging infectious diseases in recent decades have caused widespread social and economic disruptions in the global economy. Various guidelines for pandemic influenza planning are based upon traditional infection control, best practice and evidence. This paper describes the development of an early warning system for detecting disease outbreaks in the urban setting of Hong Kong, using 216 confirmed cases of H1N1 influenza from 1 May 2009 to 20 June 2009. The prediction model uses two variables – daily influenza cases and population numbers – as input to the spatio-temporal and stochastic SEIR model to forecast impending disease cases. The fairly encouraging forecast accuracy metrics for the 1- and 2- day advance prediction suggest that the number of impending cases could be estimated with some degree of certainty. Much like a weather forecast system, the procedure combines technical and scientific skills using empirical data but the interpretation requires experience and intuitive reasoning.

Keywords

Disease modeling; infectious disease; H1N1; early warning; Hong Kong

1 Introduction

The outbreaks of new and emerging infectious diseases like the severe acute respiratory syndrome (SARS) (CDC 2012a), the H5N1 avian influenza (CDC 2013), H1N1 swine flu (CDC 2012b), H7N9 avian influenza (CDC 2013), and the recent Ebola outbreaks (Rivers *et al.* 2014) have brought new challenges and heighten health awareness around the world. The accelerated mutation of viruses and increased risks of animal-to-human transmission have made pandemic preparedness a top public health priority in many countries. Because the occurrence of a pandemic could cause widespread social and economic disruptions (Danziger 1994, Meltzer *et al.* 1999), some researchers have asserted that pandemic prediction and simulation by modeling past events are important means of informing timely intervention to prevent the spread of diseases through a population (Yasuda *et al*. 2008, Cooley *et al*. 2011, Wu and Cowling 2011).

Conventional models on disease transmission dynamics are deficient in considering the heterogeneous nature of a population and place-specific concerns (Dushoff and Levin 1995, Sattenspiel and Dietzv 1995, Small and Tse 2005). In recent decades, there has been more attention on the integration of spatial and temporal approaches to develop better surveillance and modeling methods. Past research that examined the diffusion patterns of infectious diseases have used spatio-temporal and geostatistical methods to identify disease clusters (Lai *et al.* 2004, Sonesson 2007) and simulate disease spread based on characteristics and activities of individuals (Brouwers *et al.* 2009, Lee and Wong 2011). An increased interest in the development of spatio-temporal forecast systems is also evident in recent years. Boni *et al.* (2009) put forward a mathematical construct founded on the age-structured gravity and the

traditional Susceptible-Exposed-Infectious-Recovered (SEIR) models to assess the progression of H1N1 influenza in Vietnam. Hooten *et al.* (2010) proposed the Susceptible-Infectious-Recovered-Susceptible (SIRS) framework to emulate the dispersal of influenza virus in the United States. In a larger scale study for selected countries around the world, Gonzalez-Parra *et al.* (2011) demonstrated that the inclusion of spatio-temporal characteristics in the classical SEIR model could improve the ability to simulate multiple waves of H1N1 influenza outbreaks.

Despite their threats to public health, many infectious diseases can be contained with effective infection control measures such as early detection and sensible segregation (Connolly *et al.* 2004, Lau *et al.* 2004a, Lau *et al.* 2004b, WHO 2003). The International Health Regulations in 1969, and later revised in 2005, require disease reporting to the World Health Organization (WHO) to help with its global surveillance and advisory role in issuing travel advisories that restrain spatial interaction in and out of infected areas to keep the disease from spreading further. Although surveillance provides an opportunity for early intervention at the population level, it has remained weak in forecasting the levels and trends of disease spread for increased epidemic preparedness in public health (Mayer *et al.* 2001). We believe that an early identification of high risk locations and a spatial understanding of the disease transmission dynamics are instrumental in devising effective control mechanisms. With the availability of more detailed real-time surveillance data at the national and local levels, we employed the geographic information systems (GIS) technology to explore individual-based computational models to examine interactions between infectious agents and their hosts, disease spread, prediction systems, and response strategies.

This paper describes the development of an early warning system for detecting disease outbreaks in the urban setting of Hong Kong, one of the most densely populated cities in the world. It also demonstrates the performance of a spatio-temporal and stochastic SEIR (stsSEIR) model based on the best available H1N1 influenza surveillance data. The stsSEIR model illustrates how an influenza outbreak might travel through a city by means of computer simulations and modeling using a GIS. It also shows how policymakers could make use of the information to decide which local areas should receive more resources for containing or minimizing impacts of a communicable disease.

2 Data and Methods

2.1 Data

We obtained from the Hospital Authority and the Department of Health a total of 216 confirmed cases of H1N1 influenza from 1st May 2009 to 20th June 2009. The patient records included the following data: residential address, onset date of symptom, date of diagnosis, health conditions, and demographic characteristics. In the development of an early warning system, our focus was on the earlier cases before a local disease outbreak that can collectively signal an alarm for possible local outbreaks in the next few days. In addition, the most recent population census data (C&SD 2006) and related geographic data on land uses (CUHK 2001), and building locations were obtained from relevant government or corporate sources to account for social mixing in an urban environment of Hong Kong.

2.2 Model Development

The stsSEIR model is spatio-temporal in nature whereby the spatial constituent is the territory of Hong Kong and the temporal factor is the daily occurrences of H1N1. The spatial data of Hong Kong was first rasterized into 864 units of 1000mx1000m square cells which excluded non-populated areas comprising primarily of country parks and restricted zones (Figure 1). This resolution was set because previous research have found that the general activity area of an individual ranges between 800 to 1000 meters (Forsyth 2006). Moreover, finer resolution does not necessarily help in making district-level policy decision (Lee and Wong 2011). Each cell was assigned a discrete value using the finite-difference method which is a major method for data discretization in complex spatial models such as those used in weather predictions (Hong Kong Observatory 2012). This discretization of data was necessary to improve computational efficiency and facilitate manipulation of spatial interaction by means of the GIS overlay technique.

- Insert Figure 1 -

The disease modeling algorithm is a modified version of the SEIR model (Anderson and May 1991). We integrated into the model diffusion heuristics and spatial databases on a GIS platform. These modifications allowed for more weights to be assigned to potential locations of social mixing (such as transport interchanges, amenity centers, health care centers, and populated areas). A related study based on similar data (Lai *et al.* 2014) found that population density is useful and the only significant variable in improving the forecast accuracy. However, it should be noted that the model has provision to accommodate additional variables by consolidating relevant factors into a

combined weighted score at each cell location. Additionally, individual-based and stochastic processes by incorporating population movement scenarios were implemented to simulate the spread of disease across space and time (Meng *et al.* 2005).

Climate and weather factors were not considered in the diffusion model because the former is more about long-term change in environmental conditions affecting disease ecology whereas the latter is likely to exert little effect within the period of disease outbreak. More explicitly, our disease models involve in-situ measurements of infected cases at the geographic scale of a city where climatic variation would be of limited capacity. The minute difference in weather variables (with daily ranges of temperature and humidity at 2.4^{o}C and 21% respectively) at the early stage of a disease outbreak would hardly play a noticeable role in disease diffusion. The following is a list of equations and parameters used in the study.

$$\frac{dE}{dt} = \beta \frac{SI}{N} \underbrace{-E_o}_{\text{Exported, cross-cell transmission}} \underbrace{+E_i}_{\text{Imported, cross-cell transmission}} - \sigma E \qquad \text{Equation 1}$$

$$\frac{dS}{dt} = -\beta \frac{SI}{N}$$

$$\frac{dI}{dt} = +\sigma E \ - \nu I$$

$$\frac{dR}{dt} = +\nu I$$

$$S + E + I + R = N$$

where the variables *S*, *E*, *I*, and *R* represent the number of suspected, exposed, infected, and removed individuals respectively within the total population *N*; ν, σ, β, and *R* represent respectively infectious period, latent period, effective contact rate, and

reproduction number. These values were derived from the basic reproduction number R_0, which was estimated by health officials at the time of the outbreak. Equation 1 is the key equation which contains two additional terms of E_o and E_i representing the exported and imported cross-cell transmission respectively.

2.2.1 The basic variables

The basic reproduction number (R_0) of an infectious disease is the number of cases that a single case would generate on average over the duration of an infectious period (Cowling *et al*. 2010). Because a larger R_0 value signals the harder it is to control an infection, this metric is useful in informing whether or not and how rapid an infectious disease can spread through a population. The proposed disease modeling was built on official R_0 values of the 2009 pandemic H1N1 notifications as computed by Cowling *et al.* (2010). This is the only non-constant parameter that fluctuates in accordance with the daily trends of a disease outbreak. Our model is flexible and can accommodate, on a daily basis, variable R_0 values ranging between 0.75 on day 1 and the peak value of 1.55 on day 10. In addition to R_0, the other required parameters of infectious period (v) and latent period (σ) are often determined at the onset of an outbreak by medical experts familiar with characteristics of the infectious disease. Finally, the effective contact rate (β) can be deduced from the aforementioned parameters (Kwong 2010).

At the beginning of a disease outbreak when $t = 0$, the numbers of S (susceptible), E (exposed), I (infected), R (removed) individuals for each lattice cell were determined by the stsSEIR model based on the above epidemiological parameters. Spatial heterogeneity is embedded in the grid structure wherein each grid cell carries a different set of S, E, I,

and *R* individuals based on their residential addresses. For each grid cell, the exposed (*E*) population at time (t) was computed using the set of equations from the stsSEIR approach described above, i.e., as a function of the susceptible (*S*) and infected (*I*) population within the total population (*N*) and considering the effective contact rate (*β)* but excluding population that became infectious. The differential equations for other compartments of susceptible (*S*), infected (*I*), and removed (*R*) population were derived accordingly and the sum of the *S, E, I,* and *R* components is held constant as the total population (*N*) within each grid cell.

2.2.2 Within-cell and cross-cell transmissions

Daily influenza cases were key inputs to the stsSEIR model over and above data about the population and other environmental attributes. The SEIR model was used to redistribute daily cases to relevant cells to minic disease transmission across the geographic space. As population is the denominator in disease modeling and burden estimation (Tatem *et al.* 2011), it was included in each cell to simulate the transmission of influenza within and across cells. The number of infected individuals for each grid cell on subsequent days was estimated by considering two factors: (i) spatial heterogeneity for within-cell transmission and (ii) spatial proximity for cross-cell transmission (Figure 2). Spatial heterogeneity considers the impact of localised SEIR component values on the disease contact rate. Here, larger susceptible (*S*), infected (*I*), and exposed (*E*) population expect to increase not only the chance of within-cell disease transmission but also disease transmission to neighbouring cells albeit at varying degrees. Spatial proximity is governed by three kinds of cell in the modelling of cross-cell transmission: (a) pivot – a moving cell that visits every cell in the grid, (b)

neighbouring – grid cells surrounding the pivot cell, and (c) other – grid cells selected by random processes (Figure 3). Besides within-cell transmission, the population in the pivot cell has a higher probability of being infected by diseased individuals living in neighbouring cells of close proximity and vice versa. Infectivity beyond the living environment (e.g., between co-workers, school mates, or people on the streets) is also likely, and this is modelled through a weighted random process.

--- Insert Figures 2 and 3 ---

Cross-cell transmission for each pivot cell has two constituents: exported and imported transmission (Equation 1). For exported transmission, the stsSEIR model assumes individuals in the pivot cell to engage in two levels of interaction: a higher level with residents in its neighbouring cells and a lower level with those from other cells (Figure 3). Based on Tobler's (1970) first law of Geography stating that "everything is related to everything else, but near things are more related than distant things", a higher probability shall be assigned to neighbouring cells closer to a pivot or an infected cell than the more distant random cells. The same reason is made for imported transmission in which a higher probability of interaction with a pivot cell comes from neighbouring as opposed to other random cells.

2.2.3 Control parameters

A stochastic mechanism is introduced in the stsSEIR model in estimating both exported and imported cross-cell transmission with neighbouring and other cells. The model allows a certain proportion of the exposed (*E*) population within a pivot cell to contract disease from its neighbouring and other cells randomly chosen by the system in

consideration of the social mixing potential. The probability of within-cell transmission, cross-cell transmission to neighbouring cells, and cross-cell transmission to other random cells can be adjusted according to advice by medical experts given their knowledge about a disease. As the total number of infected cases is controlled by the reproduction number R_t, the proportion of one type of transmission will affect the other two.

In short and knowing that each center cell can have up to 8 first-order nearest neighbors, the neighborhood analysis was employed to simulate disease transmission between cells in two ways: (i) weighted nearest neighborhood transmission; and (ii) weighted random cross-cell transmission. These cells were determined by a weight table derived by overlaying population and place specific features. The proportions of neighborhood versus random transmissions were set as 30% and 50% respectively of the total amount of transmission. The remaining 20% was automatically allocated to within cell transmission. The full set of percentages was chosen by a team of infectious disease experts based on their clinical experience. When the model is applied in a real situation, the model operator should seek expert opinion about the percentages because diseases have varying latent and infectious periods to exhibit different characteristics of spread.

Although the stsSEIR model is not perfect, it is capable of accommodating both temporal and spatial variability not possible with deterministic models. Instead of returning a single value for the whole of Hong Kong for each model run, our stsSEIR model yields different values for each populated cell which can be further aggregated by

definable neighborhoods or administrative units for decision making at the district or community levels.

2.3 Model Validation

The 1-7 day time-series forecasts of infected cases for the whole of Hong Kong were estimated using the stsSEIR model described above. The model was developed using actual H1N1 data in the months of May and June 2009 in which data reported in May were used to predict outcomes in June. The 1-7 day forecast results were computed successively for the validation period (i.e., 1-7, 2-8, 3-9, … until 14-20 June) to allow for comparison against the actual incident data over the same period. Input to the forecast model for each set of 1-7 day forecast was continuously updated using actual incident data before the days of extended forecast to simulate real-time updates as more incident cases were notified. A time-series comparison of 1-2 day and 6-7 day forecast accuracy based on daily counts was plotted to illustrate the magnitude difference between immediate versus extended forecasts (see also Gonzalez-Parra *et al.* 2011).

Three additional metrics were computed for 1- to 7-day advance forecasts to assess the overall prediction accuracy of the stsSEIR model for early warning purposes: R^2, average absolute error (AAE), and maximum absolute error (MAE). Generally speaking, better prediction accuracy is indicated by a larger R^2 and smaller values for both AAE and MAE. The performance of the stsSEIR model against two common models by simple linear regression (LR) and autoregressive integrated moving average (ARIMA) methods was also conducted. The spatial version of LR and ARIMA models were not considered in this study because of technical limitation due to low disease counts in the

early stages of an outbreak. An exceeding large number of grid cells with a zero value violates the model assumption and renders the spatial version of LR and ARIMA models and their statistical inference invalid. As both LR and ARIMA spatial models cannot handle the zero-inflated situation, the model performance was assessed using only forecast errors (i.e., total of all cell count errors) for the whole of Hong Kong.

The disease diffusion pattern was examined using error mapping (Berry 2007). Time-series maps of 1-day ahead forecast were aligned with the corresponding maps of actual disease counts on the same day. A similarity map was compiled to display variation between actual and forecast maps by highlighting the whereabouts of very similar or dissimilar locations. The map uses beige and pale green to indicate cell locations of fairly similar counts (i.e., a difference of zero to one count) and purple or red to indicate dissimilar areas (with a larger difference of 3 or more counts).

3 Results

3.1 Predicting Disease Outbreaks

Two sets of line graphs were used to illustrate the prediction difference between immediate (1-2 days) and extended (6-7 days) forecasts against the actual or observed daily counts of disease incidents (Figures 4a and 4b respectively). It is clear from the linear plots of actual counts in Figure 4 that a small surge in disease counts or the first peak occurred on 11 June (day 11) and a sharp increase in infection or the second peak took place on 17-18 June (days 17 and 18). These increases could be attributed to the presence of small-scale local outbreaks. Comparing the linear plots of forecast values in

Figures 4a and 4b, we noted that the immediate (1-2 day) forecasts were more sensitive in detecting changes in disease counts whereas the extended (6-7 day) forecasts exhibited a smoothing effect that concealed minor fluctuations in values along the temporal scale. Besides, only the 1-day forecast was able to signal the considerable increase in infection or the second peak on days 17 and 18 based on recent cases.

- Insert Figure 4 -

R^2 values as well as AAE and MAE were computed for each prediction to offer more objective assessment of the forecast results. Very high R^2 values of 0.78 and 0.8 were obtained respectively for the 1- and 2- day advance forecasts (Table 1) indicating a high correlation between forecast and actual disease counts. R^2 values for the remaining 3- to 7- day forecasts were comparatively lower but remained acceptable within the range of 0.30 to 0.49. Incidentally, both the AAE and MAE registered the smallest values for 1-day forecast with those of 2-day forecast lagging slightly behind. Table 1 also shows that our stsSEIR model tended to over predict disease occurrences, yielding AAE and MAE ranging from 6 to 9 cases and from 19 to 33 cases respectively. Besides, the results also showed that New Territories had the lowest average R^2, whereas Hong Kong Island had the highest average R^2. When predictions were subdivided by East and West New Territories, the average R^2 values went below the prediction levels for the whole of New Territories, signaling increased prediction uncertainty at each level of spatial disaggregation.

- Insert Table 1 -

Table 2 shows performance comparison of the stsSEIR model against those of the LR and ARIMA models. The stsSEIR yields better R^2 and MAE values when compared

with both the LR and ARIMA models for the 1- and 2- day advance forecasts. However and except for the metrics of MAE, the stsSEIR does not show consistent performance for extended forecasts beyond two days. Based on the R^2 measure, it appears that the stsSEIR has better accuracy for immediate forecasts but the ARIMA excels in extended forecast albeit with larger MAEs.

- Insert Table 2 -

Besides model validation based on aggregated results described above, the stsSEIR model appears capable of forecasting spatio-temporal patterns of disease distribution. Table 3 shows a frequency table of error counts for 1-day advance forecasts from 1 June to 17 June, which is a more precise quantitative description of the temporal error. It is clear that the forecasts were very good because over 94% of the 864 grid cells were error free. The forecast on 17 June was the worst among the 17 days with 52 (6%) grid cells in disagreement and an MAE of 3 cases. The similarity maps (rightmost column of Figures 5a-5c) display the spatio-temporal variation between actual counts and forecast results (first two columns of Figures 5a-5c respectively). The overall predictability of the stsSEIR model on 17 June was 94% as shown by a large proportion of similar cells. Indeed, forecast maps from 1-17 June show comparable patterns with dispersed locations containing disease cases. Moreover, the error maps were very silent before 11 June after which both the number and magnitude of error cells continued to increase, indicating increased spatial uncertainty over time.

- Insert Table 3 and Figure 5 -

4 Discussion

The modeling approach adopted for this study is raster GIS processing based on grid cells of 1000mx1000m in size. This process of discretization is important for two reasons. Firstly, it facilitates the modeling of a large and continuous geographic region by splitting it into smaller units of uniform resolution at a suitable scale for computational efficiency. Secondly, it enables the reconstitution of geographic areas (based on neighborhoods or districts or other appropriate administrative units, see Figure 1) through aggregation of cells to yield health events and health characteristics of populations for decision making. Although the approach does not evade entirely the modifiable area unit problem (Rushton 1998), it does offer the flexibility of exploring changing effects of administrative units on public health decision making. This study also found that the spatial resolution at 1000mx1000m not only affords both computational efficiency but also provides the best forecast accuracy. More refined grid cell resolution of 500mx500m or 200mx200m might improve the visual presentation but a higher resolution would mask health effects because of more data scattering and insufficient explanatory powers with too few cases to report in the earlier phases of a disease outbreak. Although not examined in this paper, it should be noted that inappropriate spatial aggregation may impair decision making or cause delays in the deployment of very specific disease control measures whereas proper data aggregation may reduce false alarms resulted from unsuitably detailed spatial scales (Lee and Wong 2011).

A model with fairly high predictive power for an impending disease outbreak would provide public health authorities advance notice for mobilizing actions to prevent

further spread of the disease agent. Even if a disease outbreak could not be prevented, the available resources could be better allocated among needy regions based on the model forecasts. The two sets of line graphs as shown in Figures 4a and 4b complement each other in the decision process – the former seems more sensitive to recent disease occurrences but the latter compromises sensitivity to reveal a more general trend. For example, the small surge in disease counts on day 11 might not warrant issuing an alert that might trigger unnecessary public panic. By considering the highly fluctuated (1-2 day immediate) versus the smoothed (6-7 day extended) forecast results along with values of the accuracy metrics (R^2, AAE and MAE), public health authorities can determine with some degrees of certainty whether the predicted increase in infection is above a reasonable threshold and whether it is the appropriate time to issue health warnings.

The series of maps showing the spatial distribution of both actual and forecast disease locations are effective in revealing potential problem areas needing attention (Figures 5). Spatially close neighbors may not display significant interaction. The error maps (rightmost column of Figure 5) that show similarity and differences between forecast and actual disease counts are useful in summarizing locations of agreement and disagreement. Their combined use enables visual and quantitative assessments of the spatio-temporal disease interaction. Although the predicted disease clusters were generally more compact (middle column of Figure 5) whereas the actual disease clusters (leftmost column of Figure 5) were more widespread, the difference would not affect policy makers in targeting intervention measures as long as the locations of the clusters could be reasonably predicted. For example, the error maps indicated increasing prediction uncertainty both in magnitude and spatial location after 11 June. This increase

in error could be related to the small disease outbreak on 11 June which had a greater impact on densely populated areas, such as the middle of Kowloon which was correctly predicted by the stsSEIR model even though not at the exact cell locations. We would attribute the uncertainty to random effects (perhaps not properly captured by the model) but also to intervention or control measures issued after the outbreak (not considered in the model).

The integration of spatial and time-series analysis techniques seemed to avoid either missing spatial or temporal changes were the techniques applied in isolation from one another, such as relying solely on spatial linear regression or typical time series analyses. This observation was supported by results of our model validation revealing that the stsSEIR outperformed two common models in the 1- and 2- day advance forecasts (see Table 2). Moreover, the MAE provides critical information about the minimum resource required to cope with the worst-case scenario for a surveillance/monitoring system related to life-threatening infectious diseases. The consistent performance of the stsSEIR on MAE for forecasts beyond two days is very encouraging and implies that the worst-case forecast can be improved. However, the underlying condition for successful application of the stsSEIR model is rapid and systematic disease surveillance to track accurately the ongoing development of disease occurrence and disease potential. A future development direction is to compare the stsSEIR with other successful spatio-temporal models, such as the Bayesian maximum entropy model (Kolovos *et al.* 2013), in search of optimal models for select diseases.

One limitation of the stsSEIR model is its inability to predict sudden upsurge of disease cases such as that in the 2003 SARS outbreak (Nelson 2007). The

neighborhood concept for within and cross-cell infection was formulated on the principle of distance decay in which near things are more related than distant things (Tobler 1970) in conjunction with the city's population profile. But, highly infectious cases originated from visitors of other cities or countries are not fully accounted for in the stsSEIR model except through the stochastic events. Further modification of the model is required to cater for such external and random occurrences that yield contact patterns different from those of ordinary work/school travels.

It appears that the stsSEIR would be applicable to diseases similar to H1N1, such as H7N9 but not HIV/AIDS because the means of disease transmission is different between the two diseases. Moreover, as population density is an important element for developing the stsSEIR, it is expected that the model would be applicable in urban settings similar to Hong Kong but not necessarily to rural areas with a low population density and a different mobility pattern unless with modification. This reasoning was based on decreasing R^2 values from densely populated (Kowloon) to less populated (New Territories) regions because few disease cases rendered the stsSEIR model less effective, as observed from this study.

5 Conclusions

Overall, our stsSEIR model successfully predicted the H1N1 disease outbreaks in the early stages of disease transmission. The high R^2 values and reasonably low MAEs for the 1- and 2- day advance forecasts suggest that the number of infected cases could be predicted with some degree of accuracy that would not be too far from the actual value. This finding is important because public health officials would be more confident in

estimating resources to cope with the worst situation by knowing the largest possible number of infected patients (forecast value plus maximum error). The model developed by this research is valuable because the existing three-level response system to cope with disease outbreaks in Hong Kong is deficient (CHP 2012). Each level of response warrants a given set of public health actions corresponding to risk-graded epidemiological scenarios. Although the actions are designed to comply with the WHO's guidelines for pandemic influenza planning, they would be made only after the epidemiological scenarios had occurred. The stsSEIR model developed in this research outperforms the existing response system in the temporal dimension because it can predict disease outbreaks in advance to forewarn government officials before they occur. Although the precision of spatial forecast still needs further investigation, the ability to predict and visualize potential disease clusters can be an effective means for policy makers to operationalize abstract numeric predictions into location specific intervention policies. By identifying potential locations for disease outbreak, policy makers can plan ahead to mobilize community centers or markets in specific neighborhoods to provide free hygiene packages containing hand sanitizers and masks, as well as bins to allow for proper disposal of waste tissues. In locations with potential for a large outbreak, contingency plans for educational continuity or centralized care provision can be made for possible closing of schools or elderly homes in specific neighborhoods based on their demographic characteristics.

A multi-staged early warning approach similar to the routinely used typhoon warning system (http://www.hko.gov.hk/wxinfo/currwx/tc.htm) could be established for infectious disease epidemics such that response plans can be gradually ramped up as forecast certainty increases. This approach would give public health officials as much

advance notice as possible about the likelihood of a disease outbreak in a particular location to weigh the costs of response actions against the risks posed to the public. Ultimately, an early warning system can facilitate decision making by relevant national and local-level agencies and to enable at-risk individuals and communal groups to take precautionary actions against an impending disease outbreak. However, the model's predictability depends on the ability to provide accurate counts and whereabouts of the infection. This requirement suggests the need to operationalize field epidemiological data collection and reporting systems to offer rapid and near real-time disease surveillance and seroprevalence data. Careful coordination and cooperation among relevant parties, along with the advent of mobile computing and anonymous location tracking, can better enable on-target control strategies and intervention measures.

6. Acknowledgements

We are grateful to the following government departments of the Hong Kong Special Administrative Region for data access: Census and Statistics Department, Hospital Authority, Lands Department, and Planning Department. This paper is the result of research collaboration between Princess Margaret Hospital, Hospital Authority, and Department of Geography at the University of Hong Kong. The project is funded by the Research Fund for the Control of Infectious Diseases administered by the Food and Health Bureau and the Hong Kong SAR Government.

7 References

Anderson, R.M. and May, R.M., 1991. *Infectious Diseases of Humans: Dynamics and Control*. Oxford: Oxford University Press.

Berry, J.K., 2007. *Topic 16: Characterizing Patterns and Relationships. In: Beyond Mapping III* [online]. Berry & Associates // Spatial Information Systems, Inc. Available from: www.innovativegis.com/basis/MapAnalysis/Topic16/ Topic16.htm [Accessed 16 June 2013].

Boni, M.F., 2009. Modelling the progression of pandemic influenza A (H1N1) in Vietnam and the opportunities for reassortment with other influenza viruses. *BMC Medicine*, 7, 43.

Brouwers, L., *et al.*, 2009. *MicroSim: Modeling the Swedish Population* [online]. Cornell University Library. Available from: arxiv.org/ftp/arxiv/papers/0902/0902.0901.pdf [Accessed 19 October 2009]

C&SD, 2013. *Statistics on Map Dashboard* [online]. Census and Statistics Department, HKSAR. Available form: www.censtatd.gov.hk/hkstat/dashboard/index_en_ 2006.html [Accessed 14 August 2013]

CDC, 2012a. *Severe Acute Respiratory Syndrome (SARS)* [online]. Centers for Disease Control and Prevention. Available from: www.cdc.gov/sars/index.html [Accessed 1 August 2012]

CDC, 2012b. *Information on Swine Influenza/Variant Influenza Viruses* [online]. Centers for Disease Control and Prevention. Available from: www.cdc.gov/flu/swineflu/ index.htm [Accessed 1 August 2012]

CDC, 2013. *Information on Avian Influenza* [online]. Centers for Disease Control and Prevention. Available from: www.cdc.gov/flu/avianflu/index.htm [Accessed 15 April 2013]

CHP, 2012. *Preparedness Plan for Influenza Pandemic, The Government of the Hong Kong Special Administrative Region* [online]. Centre for Health Protection, HKSAR. Available from: www.chp.gov.hk/files/pdf/erib_preparedness_plan_for_influenza_pandemic_2012_2012080601_en.pdf [Accessed 14 June 2013]

Connolly, M.A., *et al.*, 2004. Communicable diseases in complex emergencies: impact and challenges. *Lancet*, 364(9449), 1974-1983.

Cooley, P., *et al.*, 2011. The Role of Subway Travel in an Influenza Epidemic: A New York City Simulation. *Journal of Urban Health*, 88(5), 982-995.

Cowling, B.J., *et al.*, 2010. The effective reproduction number of pandemic influenza in Hong Kong: prospective estimation. *Epidemiology*, 21, 842-6.

CUHK. 2001. *Consultancy Study to Analyse Broad Land Use Pattern of Hong Kong* [Online]. Department of Geography, Chinese University of Hong Kong. Available from: www.pland.gov.hk/pland_en/p_study/comp_s/lup/index_e.htm [Accessed 1 August 2013]

Danziger, R., 1994. The social impact of HIV/AIDS in developing countries. *Social Science & Medicine*, 39(7), 905-917.

Dushoff, J. and Levin, S., 1995. The effects of population heterogeneity on disease invasion. *Mathematical Biosciences*, 128, 25-40.

Forsyth, A., Ed., 2006. *Twin Cities Walking Study Environment and Physical Activity: GIS Protocols* [Online], Version 3.1. Available from: designforhealth.net/ wp-content/ uploads/2012/12/MinnGIS_Ver3_1_021006Fin.pdf [Accessed 12 July 2014].

Gonzalez-Parra, G., *et al.*, 2011. Modeling the epidemic waves of AH1N1/09 influenza around the world. *Spatial and Spatio-temporal Epidemiology*, 2, 219-226.

Hong Kong Observatory, 2012. *Numerical Modelling for Weather Prediction in Hong Kong* [online]. Hong Kong Observatory. Available from: www.weather.gov.hk/wservice/tsheet/nwp.htm [Accessed 28 June 2013]

Hooten, M.B., Anderson, J., and Waller, L.A., 2010. Assessing North American influenza dynamics with a statistical SIRS model. *Spatial and Spatio-temporal Epidemiology*, 1, 177-185.

Kolovos A, *et al.* 2013. Model-driven development of covariances for spatiotemporal environmental health assessment. *Environmental Monitoring and Assessment,* 185(1), 815-831.

Kwong, K.H., 2010. *Spatio-temporal transmission modelling of an infectious disease – a case study of the 2003 SARS outbreak in Hong Kong*. PhD Thesis, The University of Hong Kong.

Lai, P.C., *et al.*, 2014. Population Factors Affecting Initial Diffusion Patterns of H1N1. *Population and Health Management*, 17(6), 390-391.

Lai, P.C., *et al.*, 2004. Understanding the spatial clustering of severe acute respiratory syndrome (SARS) in Hong Kong. *Environmental Health Perspectives*, 112(15), 1550-1556.

Lau, J.T.F., *et al.*, 2004a. Probable secondary infections in households of SARS patients in Hong Kong. *Emerging Infectious Diseases*, 10(2), 235-243.

Lau, J.T.F., *et al.*, 2004b. SARS transmission, risk factors, and prevention in Hong Kong. *Emerging Infectious Diseases*, 10(4), 587-592.

Lee, S.S. and Wong, N.S., 2011. The Clustering and transmission dynamics of pandemic influenza A (H1N1) 2009 cases in Hong Kong. *Journal of Infection*, 63, 274-280.

Lipp, E.K., Huq, A. and Colwell, R.R., 2002. Effects of global climate on infectious disease: the cholera model. *Clinical Microbiology Reviews*, 15(4), 757-770.

Mayer, T.A., *et al.*, 2001. Clinical presentation of inhalational anthrax following bioterrorism exposure, report of 2 surviving patients. *JAMA*, 286(20), 2549-53.

Meltzer, M.I., Cox, N.J., and Fukuda, K., 1999. The economic impact of pandemic influenza in the United States: priorities for intervention. *Emerging Infectious Diseases*, 5(5), 659-671.

Meng, B., *et al.*, 2005. Understanding the spatial diffusion process of severe acute respiratory syndrome in Beijing. *Public Health*, 119, 1080-1087.

Nelson, K.E., 2007. Emerging and new infectious diseases. *In*: K.E. Nelson and CFM Williams, ed. *Infectious Disease Epidemiology: Theory and Practice*. 2nd eds. Sudbury, MA.: Jones and Bartlett Publishers, 407-504.

Rivers, C.M., *et al.*, 2014. Modeling the impact of interventions on an epidemic of Ebola in Sierra Leone and Liberia. *arXiv preprint arXiv:1409.4607*. Available from: arxiv.org/ftp/arxiv/ papers/1409/1409.4607.pdf [Accessed 12 July 2014]

Rushton, G., 1998. Improving the Geographical Basis of Health Surveillance using GIS. *In*: A.G. Gatrell and M. Loytonen, ed. *GIS and Health*. Philadelphia, PA: Taylor & Francis, 63-80.

Sattenspiel, L. and Dietz, K., 1995. A structured epidemic model incorporating geographic mobility among regions. *Mathematical Biosciences*, 128, 71-91.

Small, M. and Tse, C.K., 2005. Small world and scale free model of transmission of SARS. *International Journal of Bifurcation and Chaos*, 15(5), 1745-1755.

Sonesson, C., 2007. A CUSUM framework for detection of space–time disease clusters using scan statistics. *Statistics in Medicine*, 26, 4770-4789.

Tatem, A.J., *et al.*, 2011. The effects of spatial population dataset choice on estimates of population at risk of disease. *Population Health Metrics*, 9, 4.

Tobler, W.R., 1970. A computer movie simulating urban growth in the Detroit Region. *Economic Geography*, 46, 234–240.

WHO, 2003. *Severe Acute Respiratory Syndrome (SARS): Status of the Outbreak and Lessons for the Immediate Future* [online]. World Health Organization. Available from: www.who.int/csr/media/sars_wha.pdf [Accessed 12 July 2007]

Wu, J.T.K. and Cowling, B.J., 2011. The use of mathematical models to inform influenza pandemic preparedness and response. *Experimental Biology and Medicine*, 236(8), 955-961.

Yasuda, H., *et al.*, 2008. Preparedness for the Spread of Influenza: Prohibition of Traffic, School Closure, and Vaccination of Children in the Commuter Towns of Tokyo. *Journal of Urban Health*, 85(4), 619-635.

Table captions

1 Accuracy metrics of 1-7 day forecasts

2 Comparisons of 1-7 day forecasts using different forecast methods

3 Error distribution of 1-day ahead forecasts, 1-17 June 2009

Figure captions

1 Populated land areas of Hong Kong in 2009.

Each square cell measures 1000m×1000m in size. Populated areas were identified from the 1:5000 ortho-photographs of Hong Kong by removing country parks and restricted (uninhabited) zones. The populated cells can be further aggregated into 4 geographic zones (Hong Kong Island, Kowloon, West New Territories, and East New Territories) and shaded according to population density classes based on 2006 census data.

2 Components of exposed population.

3 The mechanisms of cross-cell disease transmission.

4 Comparing actual and forecast H1N1 occurrences.

(a) 1- and 2- day actual and forecast H1N1 occurrences, 1-18 June 2009.

(b) 6- and 7- day actual and forecast H1N1 occurrences, 6-20 June 2009.

5 Similarity maps between actual and forecast H1N1 occurrences, 1-17 June 2009.

(a) Actual, forecast, and absolute error maps on 1 June, 3 June, and 5 June 2009.

Figures 5a, 5b and 5c each shows three columns of time series maps on every other day. The leftmost column displays actual disease distribution; the one in the middle shows forecast disease distribution; and the rightmost column presents the absolute errors or difference in disease counts between actual and forecast values (beige = 0 or no difference; green = 1; blue = 2; purple = 3;

= 4). A maximum difference of 2 counts was observed in the early days of disease occurrence. The forecast time series maps will change when actual counts have been updated to reflect the current position of disease incidence.

(b) Actual, forecast, and absolute error maps on 7 June, 9 June, and 11 June 2009. *A maximum difference of 2 counts was observed in the early phase when there was no significant increase in disease incidence, indicating the relative stability of the model in spatial forecast.*

(c) Actual, forecast, and absolute error maps on 13 June, 15 June, and 17 June 2009. *A maximum difference of 3 counts was observed (the respective cells were shaded purple) over the 17-day period. The sudden increase in disease incidence resulted in a larger absolute difference in the spatial forecast and alerted new locations of disease outbreak. These changes in the diseased locations signaled deviation from the model as a result of abnormal propagation of the disease. The new locations suggested potential areas needing additional medical attention to enhance preparedness and response capabilities.*

Table 1 Accuracy metrics of 1-7 day forecasts

Area	Number of days ahead forecast	R^2	Average Absolute Error (AAE)*	Average Error (AE)*	Maximum Absolute Error (MAE)*
Whole Territory	1 day	0.78	6	4	19
	2 days	0.80	7	2	24
	3 days	0.41	9	2	28
	4 days	0.30	9	2	30
	5 days	0.39	8	1	31
	6 days	0.49	8	0	30
	7 days	0.41	8	-1	33
	The cell with the best value for each accuracy metrics.				
Hong Kong Island	1 day	0.57	2	1	7
	2 days	0.71	2	1	6
	3 days	0.23	4	1	14
	4 days	0.16	4	0	13
	5 days	0.20	4	-1	11
	6 days	0.37	3	-1	10
	7 days	0.08	4	-2	12
Kowloon	1 day	0.41	3	1	14
	2 days	0.69	3	0	14
	3 days	0.36	3	0	16
	4 days	0.41	3	0	15
	5 days	0.23	3	-1	18
	6 days	0.24	3	-1	17
	7 days	0.17	4	-1	18
New Territories	1 day	0.18	3	2	7
	2 days	0.06	4	1	9
	3 days	0.02	4	2	9
	4 days	0.04	4	2	10
	5 days	0.03	4	2	9
	6 days	0.16	4	2	8
	7 days	0.09	5	2	7
East New Territories†	1 day	0.06	2	1	6
	2 days	0.16	2	1	6
	3 days	0.15	2	1	6
	4 days	0.13	2	1	6
	5 days	0.08	2	1	6
	6 days	0.20	2	1	4
	7 days	0.09	2	1	6
West New Territories†	1 day	0.10	1	1	2
	2 days	0.10	2	1	7
	3 days	0.00	2	1	6
	4 days	0.09	2	1	7
	5 days	0.05	2	1	6
	6 days	0.03	2	1	6
	7 days	0.03	2	1	5

* Based on the total number of cases within the forecast period of 1-20 June 2009.
† New Territories can be subdivided into East and West New Territories (see Figure 1)

Table 2 Comparisons of 1-7 day forecasts using different forecast methods

Number of days ahead forecast	R^2			Average Absolute Error (AAE)*			Maximum Absolute Error (MAE)*		
	stsSEIR	LR	ARIMA	stsSEIR	LR	ARIMA	stsSEIR	LR	ARIMA
1 day	**0.78**	0.52	0.44	6	**4**	5	**19**	35	37
2 days	**0.80**	0.57	0.38	7	**6**	7	**24**	36	41
3 days	0.41	**0.46**	0.16	9	**6**	8	**28**	36	43
4 days	0.30	**0.54**	0.18	9	**7**	9	**30**	34	43
5 days	0.39	**0.62**	0.22	**8**	**8**	10	**31**	35	44
6 days	0.49	0.50	**0.67**	**8**	9	11	**30**	37	43
7 days	0.41	0.25	**0.43**	**8**	10	12	**33**	43	45

* Based on the total number of cases within the forecast period of 1-20 June 2009.

stsSEIR = spatio-temporal and stochastic SEIR model; LR = simple linear regression model; ARIMA = autoregressive integrated moving average model

Table 3 Error distribution of 1-day ahead forecasts, 1-17 June 2009

Date	Error count*	Frequency	percentage	Date	Error count*	Frequency	percentage
June 1	0	858	99.3	June 10	0	854	98.8
	1	6	0.7		1	10	1.2
	2	0	0.0		2	0	0.0
	3	0	0.0		3	0	0.0
June 2	0	857	99.2	June 11	0	843	97.6
	1	7	0.8		1	21	2.4
	2	0	0.0		2	0	0.0
	3	0	0.0		3	0	0.0
June 3	0	855	99.0	June 12	0	848	98.1
	1	8	0.9		1	16	1.9
	2	1	0.1		2	0	0.0
	3	0	0.0		3	0	0.0
June 4	0	856	99.1	June 13	0	847	98
	1	7	0.8		1	16	1.9
	2	1	0.1		2	1	0.1
	3	0	0.0		3	0	0.0
June 5	0	855	99.0	June 14	0	848	98.1
	1	8	0.9		1	14	1.6
	2	1	0.1		2	2	0.2
	3	0	0.0		3	0	0.0
June 6	0	853	98.7	June 15	0	843	97.6
	1	10	1.2		1	19	2.2
	2	1	0.1		2	2	0.2
	3	0	0.0		3	0	0.0
June 7	0	853	98.7	June 16	0	832	96.3
	1	10	1.2		1	31	3.6
	2	1	0.1		2	1	0.1
	3	0	0.0		3	0	0.0
June 8	0	855	99.0	June 17	0	812	94.0
	1	8	0.9		1	43	5.0
	2	1	0.1		2	7	0.8
	3	0	0.0		3	2	0.2
June 9	0	855	99.0				
	1	9	1.0				
	2	0	0.0				

* The results are not exactly the same as those in Table 1 because of rounding process at each grid cell.

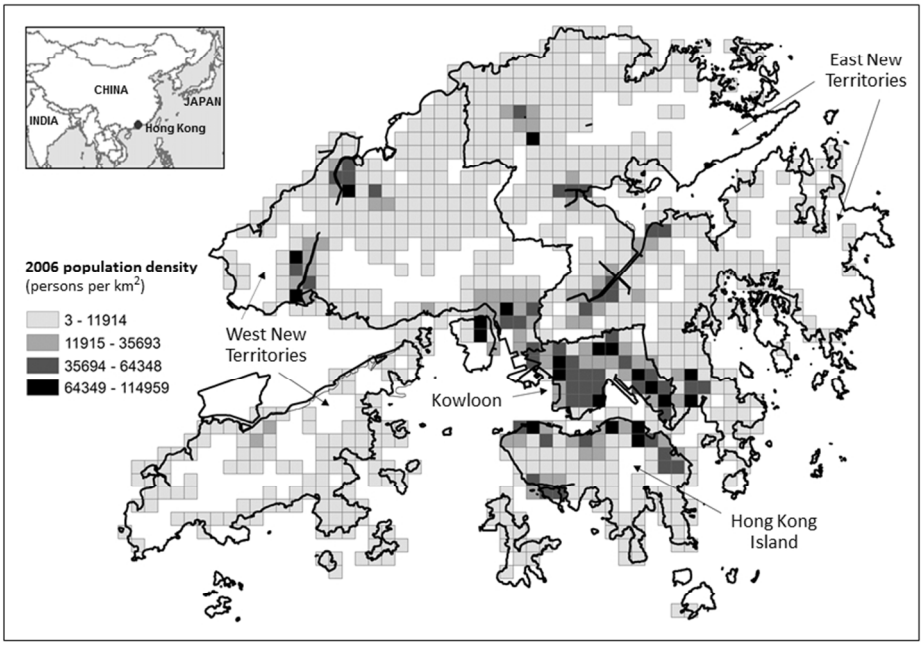

Figure 1 Populated land areas of Hong Kong in 2009.

Each square cell measures 1000mx1000m in size. Populated areas were identified from the 1:5000 ortho-photographs of Hong Kong by removing country parks and restricted (uninhabited) zones. The populated cells can be further aggregated into 4 geographic zones (Hong Kong Island, Kowloon, West New Territories, and East New Territories) and shaded according to population density classes based on 2006 census data.

254x190mm (96 x 96 DPI)

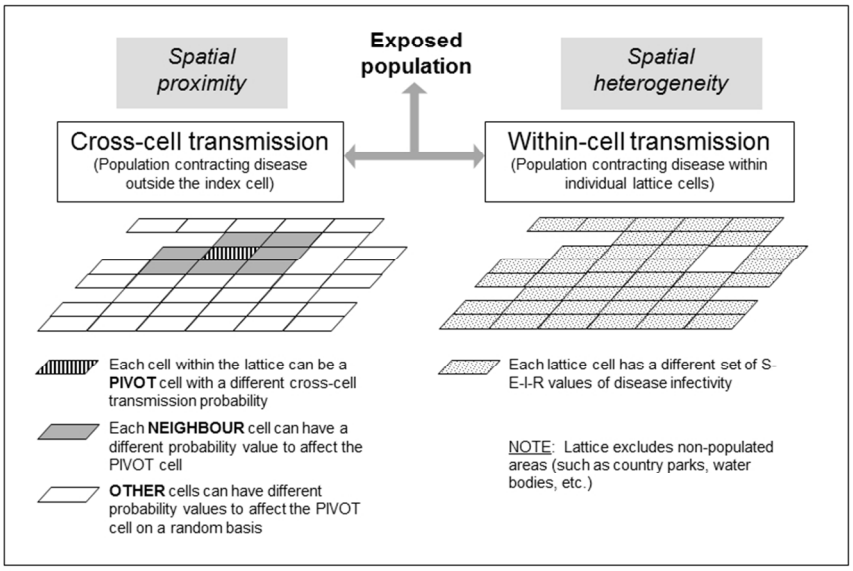

Figure 2. Components of exposed population.
254x190mm (96 x 96 DPI)

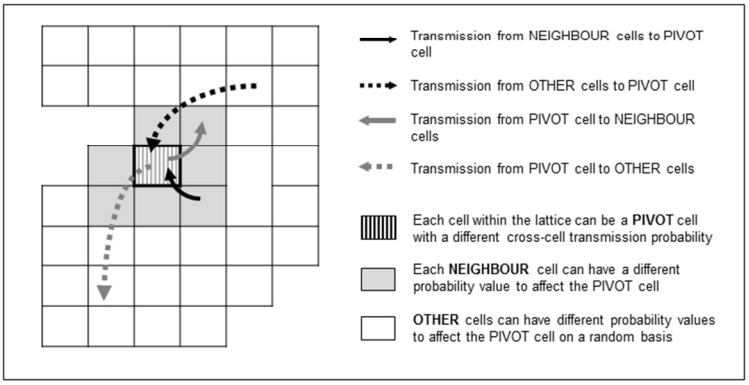

Figure 3. The mechanisms of cross-cell disease transmission.
254x190mm (96 x 96 DPI)

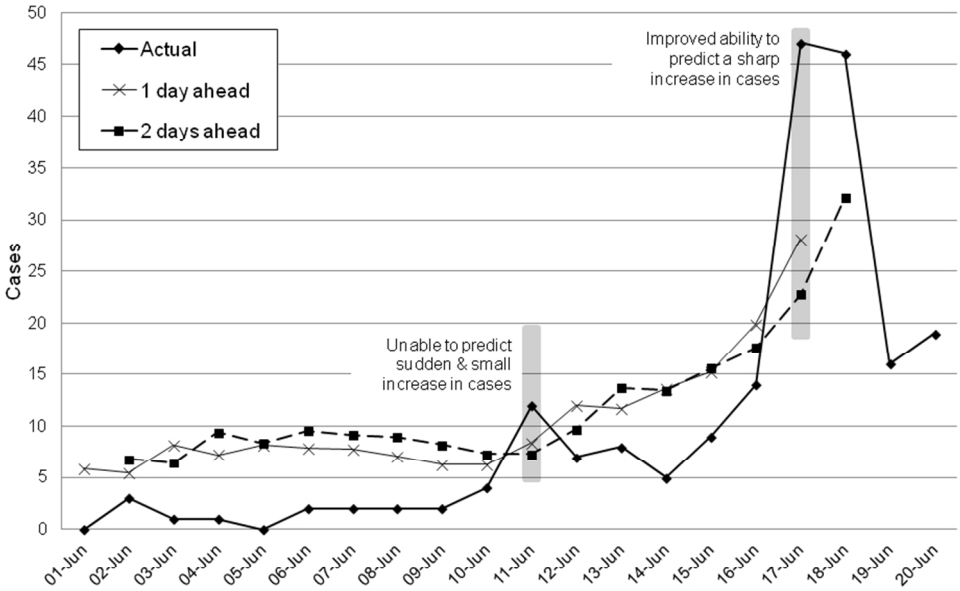

Figure 4. Comparing actual and forecast H1N1 occurrences.

(a) 1- and 2- day actual and forecast H1N1 occurrences, 1-18 June 2009
(b) 6- and 7- day actual and forecast H1N1 occurrences, 6-20 June 2009

254x190mm (96 x 96 DPI)

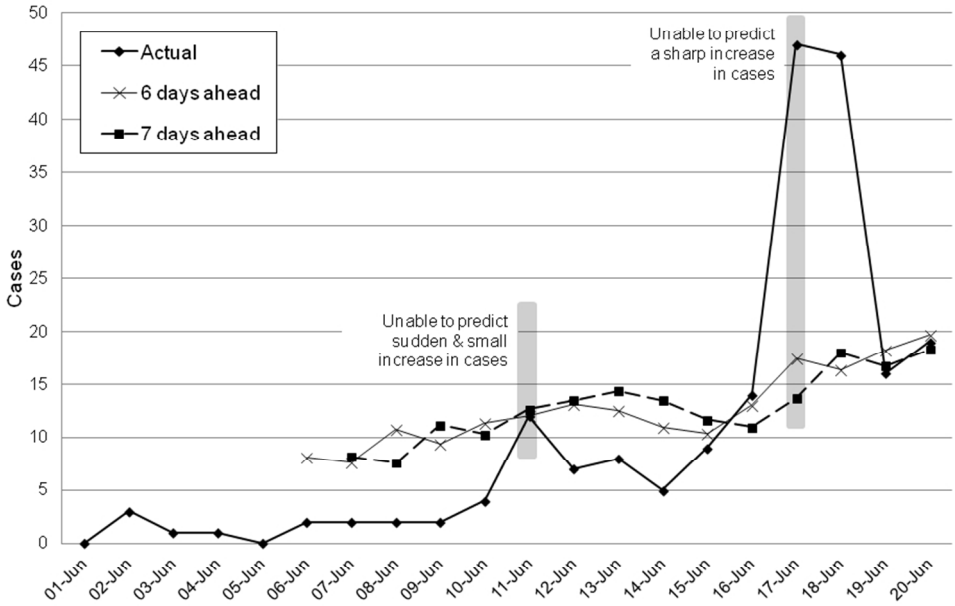

Figure 4. Comparing actual and forecast H1N1 occurrences.

(a) 1- and 2- day actual and forecast H1N1 occurrences, 1-18 June 2009
(b) 6- and 7- day actual and forecast H1N1 occurrences, 6-20 June 2009
254x190mm (96 x 96 DPI)

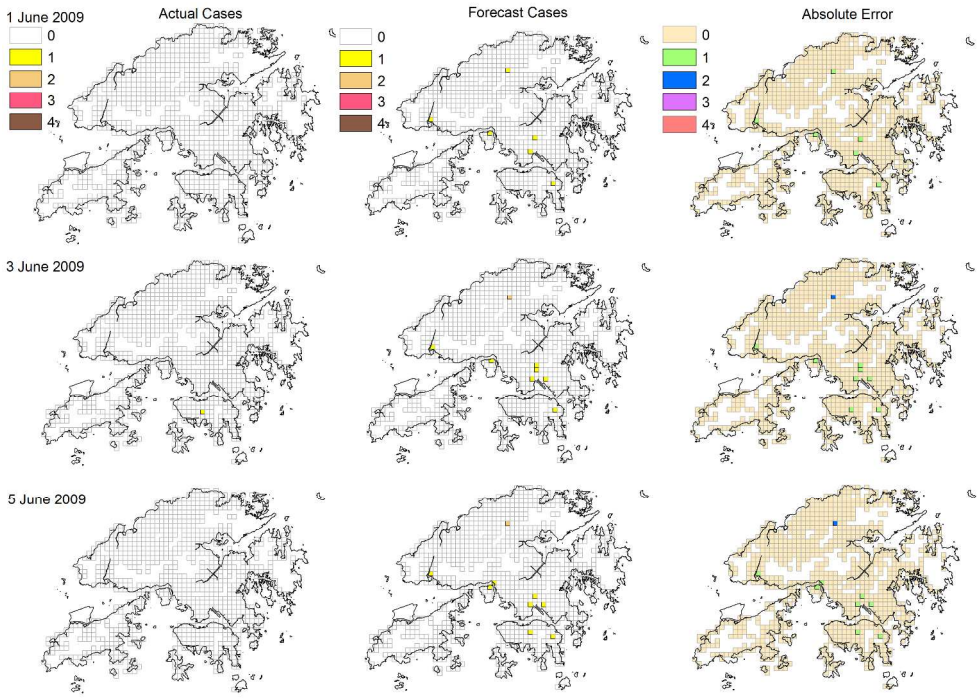

Figure 5 Similarity maps between actual and forecast H1N1 occurrences, 1-17 June 2009.

(a) Actual, forecast, and absolute error maps on 1 June, 3 June, and 5 June 2009. Figures 5a, 5b and 5c each shows three columns of time series maps on every other day. The leftmost column displays actual disease distribution; the one in the middle shows forecast disease distribution; and the rightmost column presents the absolute errors or difference in disease counts between actual and forecast values (beige = 0 or no difference; green = 1; blue = 2; purple = 3; red = 4). A maximum difference of 2 counts was observed in the early days of disease occurrence. The forecast time series maps will change when actual counts have been updated to reflect the current position of disease incidence.

297x210mm (300 x 300 DPI)

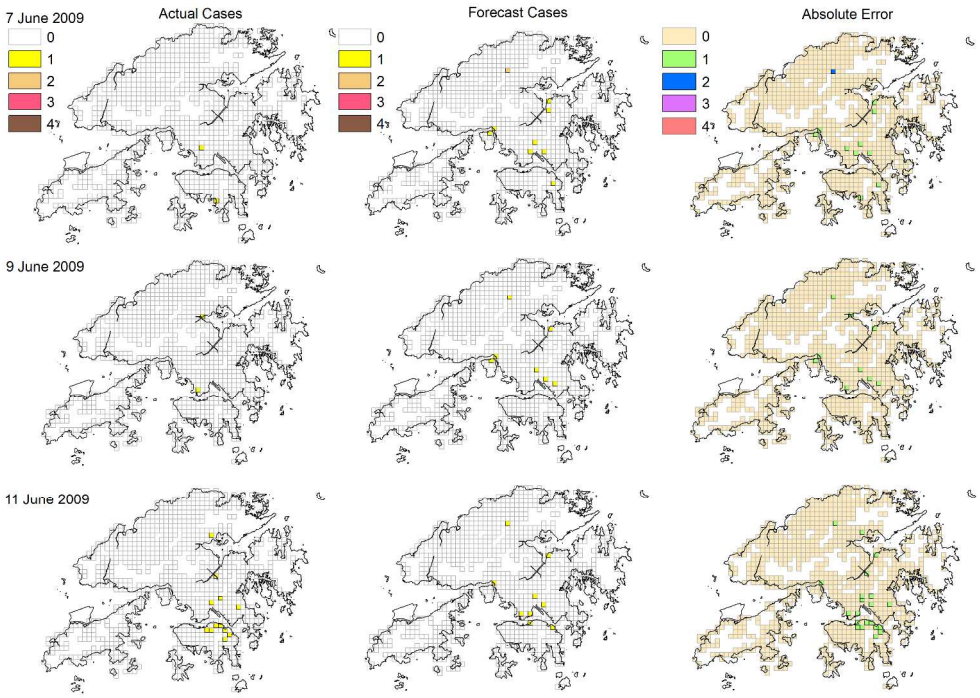

Figure 5 Similarity maps between actual and forecast H1N1 occurrences, 1-17 June 2009.

(b) Actual, forecast, and absolute error maps on 7 June, 9 June, and 11 June 2009. A maximum difference of 2 counts was observed in the early phase when there was no significant increase in disease incidence, indicating the relative stability of the model in spatial forecast.

297x210mm (300 x 300 DPI)

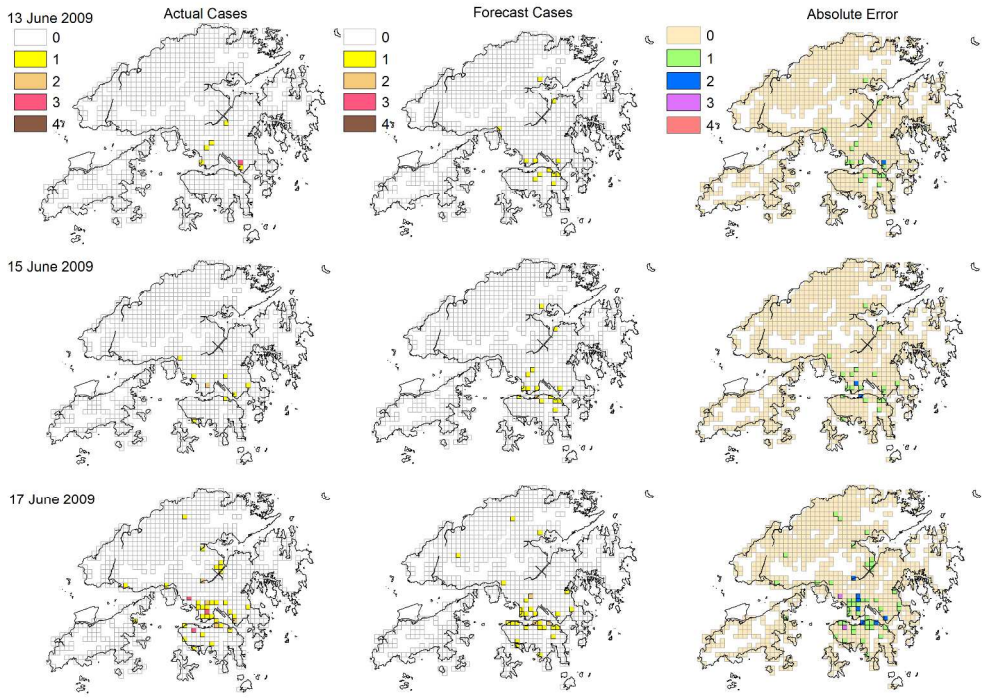

Figure 5 Similarity maps between actual and forecast H1N1 occurrences, 1-17 June 2009.

(c) Actual, forecast, and absolute error maps on 13 June, 15 June, and 17 June 2009. A maximum difference of 3 counts was observed (the respective cells were shaded purple) over the 17-day period. The sudden increase in disease incidence resulted in a larger absolute difference in the spatial forecast and alerted new locations of disease outbreak. These changes in the diseased locations signaled deviation from the model as a result of abnormal propagation of the disease. The new locations suggested potential areas needing additional medical attention to enhance preparedness and response capabilities.

297x210mm (300 x 300 DPI)