CANADIAN APPLIED
MATHEMATICS QUARTERLY
Volume 19, Number 2, Summer 2011

SOME METHODOLOGICAL ASPECTS INVOLVED IN THE STUDY BY THE BIO.DIASPORA PROJECT OF THE SPREAD OF INFECTIOUS DISEASES ALONG THE GLOBAL AIR TRANSPORTATION NETWORK

J. ARINO, W. HU, K. KHAN, D. KOSSOWSKY AND L. SANZ

ABSTRACT. The Bio.Diaspora Project studies infectious disease threats involving the rapid translocation of an infectious agent across vast distances induced by the travel of infected individuals using the global air transportation network. In this paper, the basic methodology used in the project is described.

1 Introduction The SARS epidemic of 2003 was very important in many aspects. Because of the severity of its symptoms, SARS was relatively easy to detect. Because the number of people infected was not too large, most cases were relatively well documented. The epidemic showed some of the positive and negative aspects of control measures. Although there were of course shortcomings, that it was contained in such a relatively short time is a testament to the efficacy of some of the control procedures that were put into effect.

But what is perhaps the most important teaching of the SARS epidemic are the potentially disastrous consequences of the globalization and acceleration of travel. SARS was exemplary of the ability of an emerging disease to spread very fast over large distances. The following numbers summarize easily the importance of air travel on the spatio-temporal spread of this pathogen: of the documented 137 SARS cases that are known to have crossed state boundaries, 129 traveled by plane.

Of the initiatives that followed in the wake of the SARS epidemic, two are of particular importance in the present paper. The Mathematics of Information Technology and Complex Systems (MITACS—now Mprime, Canada) put together a team of modellers and public health practitioners to work on mathematical models of infectious disease spread, which greatly facilitated communications between modellers and public health

officials in Canada. This team became the Centre for Disease Modelling, hosted by York University. Another consequence was the creation of the Bio.Diaspora Project at St Michael's Hospital, in Toronto.

The Bio.Diaspora Project studies the risks of importation of infectious agents to other health regions by means of the air transportation network. To work towards developing an understanding of the network and to be able to effectively help in public health decision processes, a set of techniques were developed or adapted. The object of this paper is to briefly present some of these techniques. Results are not presented here: the interested reader can consult some publications of the Bio.Diaspora Project such as [**7**–**11**].

2 The data The Bio.Diaspora Project focuses on air travel, although it also documents "ground conditions" in order to assess risk. In this paper, we only describe the data pertinent to air travel and human population.

Both the OAG (Official Airline Guide) and IATA (International Air Transport Association) data sets detailed below document movements out of a (mostly) common list of airports. At the time of writing, a total of 4,984 airports appeared at one time or another in one or both of the databases. Of these, a core group of a little less than 3,500 airports are active at any one time. Airports may become inactive for several reasons, and the ones that are not active throughout the span of our databases are typically very small airports or airstrips with infrequent scheduled flights.

There exists two types of letter codes to identify airports. The IATA (airport) code is a three letter identifier used for commercial purposes. ICAO (International Civil Association Organization) airport codes comprise four letters and are used for navigation. As the data we use is commercial data, we use IATA airport codes. For cities that comprises more than one airport, IATA uses metacodes that pool together the IATA codes for the airports in that city. For instance, London, UK, has IATA code LON corresponding to London Heathrow (LHR), Gatwick (LGW), Stansted (STN), Luton (LTN), City (LCY) and Biggin Hill (BQH) airports.

Two databases are specific to airports: ACI (Airport Councils International) and ICAO (International Civil Aviation Organization, a UN agency). They are not detailed here, as their content is not used in the work described in this paper.

Throughout this document, N represents the total number of airports

in the database. We refer to an arbitrary airport either using an index $i = 1, \ldots, N$ or the fictitious code XYZ.

The Official Airline Guide regroups over 1000 transporters. The database details planned regular and charter flights for the coming year. Information is relative to the companies, schedules, planned capacity and flight duration, among others. As such, it establishes the *potential traffic* for the coming year. This information is used mainly to study the network architecture. The temporal resolution of this data is the minute; the project has data for 2000-2010.

The International Air Transportation Association regroups 240 transporters accounting for 84% of flights worldwide. The data details on a monthly basis the trips taken including up to five intermediate stops. Therefore, from this data, we deduce the *effective traffic*. The project has data for 2007, 2008 and 2009.

For some applications, it is necessary to obtain information on the population that uses a given airport (see later). In order to compute this quantity, we use data from SEDAC and LandScan.

3 Characterization of the network Before discussing measures to characterize the network and study its evolution, let us give some notation for network related aspects.

3.1 Time dependence As indicated earlier, the data we have is temporal, with a resolution of one minute for OAG and one month for IATA. As a consequence, it is important to take time into account, all the more so that travel volumes vary widely depending on the period of the year; see, e.g., Figure 1.

So, in all considerations that follow, it should be understood that graphs evolve with time and the measures computed also do.

3.2 Notation Connections between airports are represented by several $N \times N$-matrices. The connection (or adjacency) matrix $\mathcal{C}(t) = [c_{ij}(t)]$ is a boolean matrix that has $c_{ij}(t) = 1$ if there is direct flight (in the OAG database) or a trip (in the IATA database) from i to j at time t, and $c_{ij}(t) = 0$ otherwise. We omit time dependence if this does not lead to confusions. To distinguish between OAG and IATA data, we denote $\mathcal{C}^O$ and $\mathcal{C}^I$, respectively, the matrices deduced using the OAG and IATA data. In this matrix as well as throughout the remainder of the text, we use the notation c_{ij} for travel from i to j, rather than the notation c_{ji} used in [**2**].

Corresponding to $\mathcal{C}$ are the matrices $\mathcal{V}$ of volumes, detailing, for any

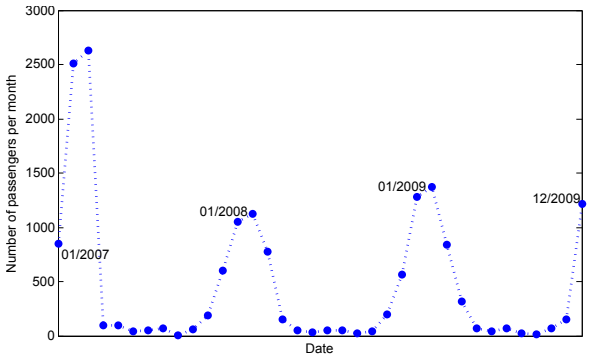

FIGURE 1: Number of trips each month originating in Winnipeg, Manitoba, Canada (YWG) and terminating in Cancun, Mexico (CUN) in the years 2007 to 2009.

pair $i, j = 1, \ldots, N$ for which $c_{ij} = 1$, the volume v_{ij} of travel between airports i and j. Note that the volume here can be either potential seats information from the OAG database, with matrix denoted $\mathcal{V}^O$, or effective trips information from the IATA database, with matrix denoted $\mathcal{V}^I$.

Another type of matrix are the connection time matrices $\mathcal{T} = [t_{ij}]$, which give the weighted median flying time between any two airports i and j. For the OAG data, the median flying time is computed as follows. If $c_{ij} = 1$, then t_{ij} is median of the flight times between i and j, weighted using the number of seats in the OAG database between i and j. If $c_{ij} = 0$, then $t_{ij} = 0$.

Corresponding to these matrices, different digraphs describe movement: $\mathcal{G}^O(t)$ is the graph obtained from the OAG data and $\mathcal{G}^I(t)$ is the graph obtained from the IATA data.

3.3 General network information The digraph $\mathcal{G}^O$ is evidently nonplanar and since $\mathcal{G}^O$ is a subgraph of $\mathcal{G}^I$, the latter is also nonplanar. For the purpose of network analysis, we generally consider the OAG digraph $\mathcal{G}^O$, since it describes the network itself and not the way the network is used. To simplify the analysis, the data is aggregated monthly: any arc that is active some time during a given month is assumed active the whole month. Several graph related measures are computed: density (ratio of arcs present in the digraph to arcs that would be present if the digraph were completely connected); strong components are iso-

lated. The following airport-related measures are computed: indegree, outdegree, degree, betweenness centrality, closeness, k-cores, cut vertices, shortest paths (geodesic –number of flights, in time, in great circle distance) and excentricity.

4 Airport catchment areas For some aspects of the work, it is necessary to have an estimate of the population that uses a given airport for its transportation needs, i.e., the population situated within the so-called *catchment area* of each airport. Because the airports are located throughout the world, it is unrealistic to gather this information manually. In order to gather this information automatically, we use a *Dirichlet tessellation* of the plane. This proceeds as follows; see, e.g., [**3**]. Let $\mathcal{P}$ be a finite set of points on a sphere, the *sources*. For each pair of points $P, Q \in \mathcal{P}$, define

$$H_{PQ} = \left\{ X : \frac{|X-P|}{\sigma(P)} \leq \frac{|X-Q|}{\sigma(Q)} \right\}$$

where $\sigma(P) > 0$, and

$$K_{PQ} := H_{PQ} \cap H_{QP} = \left\{ X : \frac{|X-P|}{\sigma(P)} = \frac{|X-Q|}{\sigma(Q)} \right\}$$

For each $P \in \mathcal{P}$, let $R_p = \bigcap_{Q \neq P} H_{PQ}$ and $R = \{R_P, P \in \mathcal{P}\}$. Then $R(\mathcal{P})$ is the Dirichlet (or weighted Voronoi) tessellation of the sphere. If the weight function $\sigma(P) = 1$ for all P, then in the plane, the regions are polygons and the result is often called a Voronoi diagram.

In the top part of Figure 2, the result of using a weight function $\sigma(P) = 1$ is shown for a region covering part of northeastern United States, Ontario and Quebec. Limitations of this weighting function is that it does not take into account the importance of the airports. Using a weight equal to the volume of trips out of an airport overemphasizes the roles of major airports, so we use Holling type 2 function of the form

$$\sigma(v_i) = v_{\max}(t) \frac{v_i(t)}{v_i(t) + v_{\text{med}}(t)},$$

where $v_{\max}(t)$, $v_{\text{med}}(t)$ and $v_i(t)$ are the volume out of the busiest airport, median volume and volume out of the airport i under consideration, respectively, from the IATA database. The tessellation is computed for every month in the database, since the relative importance of airports vary monthly.

Note that the results obtained using this method are not meant to represent the exact location where people using the airports live.

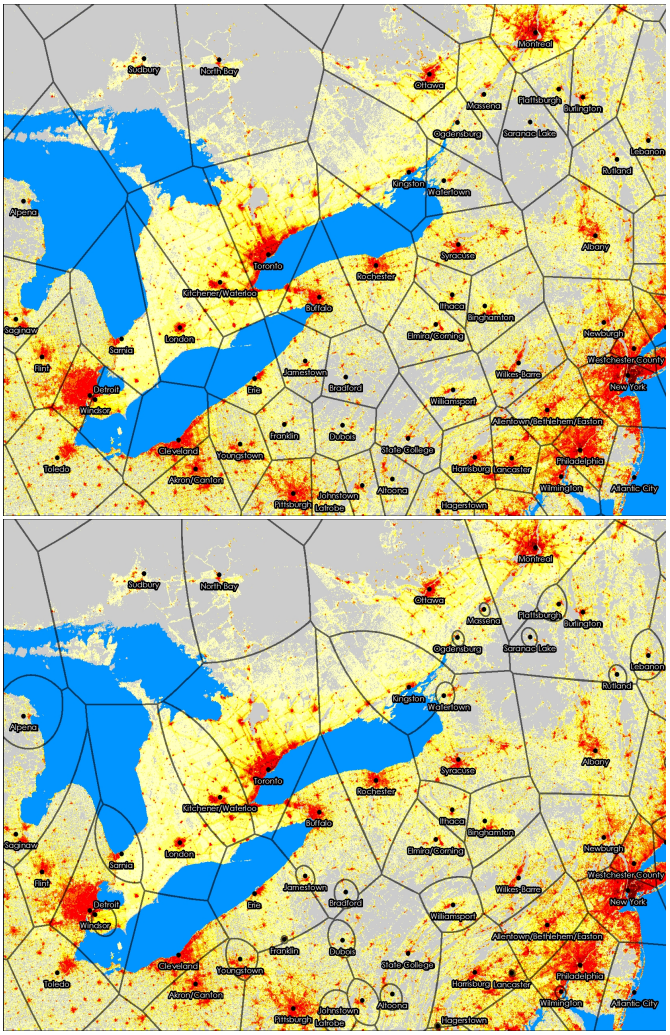

FIGURE 2: Regular Voronoi diagram and weighted Voronoi diagram obtained by considering (top) a weight function $\sigma(P) = 1$ and (bottom) a Holling type 2 weight function as explained in the text.

5 Stochastic modelling Because of the nature of the data, we consider cities with airports as the units of analysis. The population of such a city consists of the population in the catchment area of the airport serving that city. We describe the model in three steps: 1) the epidemiology in cities, 2) the description of transport and 3) the integration of both.

5.1 Description of the epidemiology in cities The model in each city $i = 1, \ldots, N$ is a classical SEIR model, which has individuals in one of the epidemiological states: susceptible, exposed, infectious and recovered, with numbers at time t in airport i denoted $S_i(t)$, $E_i(t)$, $I_i(t)$ and $R_i(t)$, respectively. When this does not lead to ambiguities, the dependence of the state variables on t is not indicated.

Prior to an epidemic event, *susceptibles* represent almost all the population. They are potentially affected by the disease, if subject to an infecting contact. Such contacts occur at the rate $S_i I_i / P_i$, where $P_i = S_i + E_i + I_i + R_i$ is the population in the airport, and result in new infections at the rate $\beta_i S_i I_i / P_i$. β_i is the disease transmission coefficient in airport i. This type of incidence is called *proportional* incidence. The disease transmission coefficient represents the probability that infection occurs, given contact. We allow it to vary from location to location, since factors such as hygiene or social distance play a role in the transmission of the disease. *Exposed* (or latent) individuals are susceptibles who have become latently infected because of an infecting contact with an infectious individual. It is assumed that patients in this state do not transmit the disease. The time spent incubating is exponentially distributed with mean $1/\varepsilon_i$ time units. *Infectious* individuals actively spread the infection through contacts with susceptible individuals. The remain infectious for an average $1/\gamma_i$ time units, with the sojourn time in the infectious class exponentially distributed. Finally, *recovered* individuals are individuals who have ceased to be infectious and are immune to reinfection (permanently in the case of an SEIR model, temporarily in the case of an SEIRS model).

Because we are interested in the course of the epidemic over a short time interval of at most a few weeks and that our focus is on the appearance of new cases in airports rather than the global course of the epidemic, we make a certain number of simplifying assumptions. Note that a nonsimplified model is also used, but for the simplicity of the exposition here, we present the simplified version.

First, we suppose that the total population in each airport is large and roughly constant, and that $P_i \approx S_i$, that is, the quantity $E_i + I_i + R_i$

is negligible compared to P_i (or S_i). This implies that proportional incidence takes the form $\beta_i I_i$. Thus, the incidence function is linear; note that this may not be true for other diseases or situations. It is also not true if the disease were considered on a longer time period, since in this case, $E_i + I_i + R_i$ might increase to such a point that S_i no longer is approximately equal to P_i. However, this assumption greatly simplifies the problem, since the system becomes independent of the population in each airport/city. Finally, we interpret the class of recovered individuals as in the first meaning it was given [**6**], in terms of *removed* individuals. Individuals are removed from the I class either by recovery or by death. Individuals in the recovered class play no role in the short term transmission of the disease, and thus we neglect this class from now on.

5.2 Transport model in the absence of infection The transport operator involves the proportion of individuals moving from city to city; denoting $p_{ij}(t)$ the proportion of individuals in city i that travel to city j, we have

$$p_{ij}(t) = \frac{v_{ij}(t)}{\sum_{j=1}^{N} v_{ij}(t)},$$

where $v_{ij}(t)$ is the volume of seats (OAG) or trips (IATA) from city i to city j in the database at time t. The quantity $p_{ij}(t)$ must then be related to the likelihood that an individual in a city does travel. In this simple model, it is assumed that all individuals in the city are equally likely to travel, and therefore, the rate at which individuals travel is equal to the ratio of the volume of travel per unit time with the total population in the city; i.e., if $P_i(t)$ is the population in city i, then the rate $m_{ij}(t)$ of movement of individuals from city i to city j is given by

$$m_{ij}(t) = \frac{\sum_{j=1}^{N} v_{ij}(t)}{P_i(t)} \frac{v_{ij}(t)}{\sum_{j=1}^{N} v_{ij}(t)} = \frac{v_{ij}(t)}{P_i(t)}.$$

This must in turn be related to the proportion of individuals in the various states. So, finally, a simple model for the rate of movement of individuals in epidemiological state $X = \{S, E, I, R\}$ from city i to city j is given by

$$m_{ij}^X(t) X_i(t) = \frac{X_i(t)}{P_i(t)} \frac{v_{ij}(t)}{P_i(t)} = \frac{v_{ij}(t)}{P_i(t)^2} X_i(t),$$

where $m_{ii}^X(t) = 0$.

5.3 General model of infection-transport The stochastic process of infection-transport can be derived in several ways, in particular, using infinitesimal probabilities. However, we show here only the most useful one for our purpose: the derivation in terms of times to transitions; see, e.g., [**1**]. Numerical simulations are indeed run using the Gillespie algorithm [**5**].

Suppose that the system is, at time t, in the state $(e, i) = (e_1, i_1, \ldots, e_N, i_N)$. Then compute

$$(1) \quad \xi_t := \sum_{j=1}^{N} \left(\varepsilon_j e_j + \left(\beta_j \frac{s_j}{P_j} + \gamma_j \right) i_j \right) + \sum_{j,k=1,\ k \neq j}^{N} \left(m_{jk}^E e_j + m_{jk}^I i_j \right),$$

the weight of possible events.

The next event then occurs at time $t + \tau_t$, where τ_t is one realization of a random variable with exponential distribution with parameter ξ_t. At time $t + \tau_t$, the transition $(e, i) \to (e', i')$ occurs, where the new state (e', i') corresponds to the following events. Note that only the variables that are modified are indicated, for simplicity.

1. A susceptible is infected in city j, i.e, $(e', i') = (\ldots, e_j + 1, \ldots)$. This occurs with probability
$$p_{(e,i)\to(e',i')} = \beta_j i_j / \xi_t.$$
2. An exposed individual in city j develops the disease (end of incubation period), i.e., $(e', i') = (\ldots, e_j - 1, i_j + 1, \ldots)$. This occurs with probability
$$p_{(e,i)\to(e',i')} = \varepsilon_j e_j / \xi_t.$$
3. An infected individual in city j recovers, i.e., $(e', i') = (\ldots, i_j - 1, r_j + 1, \ldots)$. Such an event occurs with probability
$$p_{(e,i)\to(e',i')} = \gamma_j i_j / \xi_t.$$
4. An individual currently in the incubation period travels from city j to city k (with $k \neq j$), i.e., $(e', i') = (\ldots, e_j - 1, \ldots, e_k + 1, \ldots)$, with probability
$$p_{(e,i)\to(e',i')} = m_{jk}^E e_j / \xi_t.$$
5. An infectious individual in city j travels to city k (with $k \neq j$), i.e., $(e', i') = (\ldots, i_j - 1, \ldots, i_k + 1, \ldots)$; this occurs with probability
$$p_{(e,i)\to(e',i')} = m_{jk}^I i_j / \xi_t.$$

Note that this implies that the total population is infinite.

5.4 Expected values Because the model is linear, the expected values in the model, $\langle E_j \rangle := \mathbb{E}(E_j(t))$ and $\langle I_j \rangle := \mathbb{E}(I_j(t))$, verify for all $j = 1, \ldots, N$, the deterministic model

$$\frac{d}{dt}\langle E_j \rangle = \beta_j \langle I_j \rangle - \varepsilon_j \langle E_j \rangle + \sum_{k=1}^{N} m_{kj}^{E} \langle E_k \rangle \tag{2a}$$

$$\frac{d}{dt}\langle I_j \rangle = \varepsilon_j \langle E_j \rangle - \gamma_j \langle I_j \rangle + + \sum_{k=1}^{N} m_{kj}^{I} \langle I_k \rangle, \tag{2b}$$

where $m_{kk}^{X} = -\sum_{\ell=1,\ell\neq k}^{N} m_{ik}^{X}$ for $X = \{E, I\}$. Generically, initial conditions are chosen with $\sum_{i=1}^{N}(E_i + I_i) > 1$, i.e., there is initially at least one individual who is exposed or infectious in the system of airports.

The behaviour of this type of system is well understood, in particular in the linear case here. The basic reproduction number $\mathcal{R}_0$ can be computed using the method in [**2**]. In the case of this linear system, $\mathcal{R}_0 < 1$ implies that solutions go to zero, while $\mathcal{R}_0 > 1$ implies that solutions tend to infinity (which is consistent with the stochastic system being one with infinite population).

6 Numerical simulations Travel rates are estimated using the data from IATA on trips. The population in the catchment area of each airport is estimated using the Dirichlet tessellation method.

For disease related parameters, values of the duration of the different stages are known from the literature for many diseases. In the case of an outbreak of a disease for specific parameters are not known, extensive simulations are carried out using parameters in typical ranges.

Note that because of the short time frame within which it operates, timing is essential in the present model. As a consequence, it is important to be careful when choosing values for the parameters $1/\varepsilon_i$ and $1/\gamma_i$ that represent the mean duration of stages. Suppose for example that we observe a disease in its initial stages of spread and conclude that the average incubation time is found to be on average 7 days. Inherent to the formulation of the model is that the time spent in the exposed class E_i for a given individual is an exponentially distributed random variable with mean $1/\varepsilon_i$. Considering $1/\varepsilon_i = 7$ days implies that in a cohort of individuals infected on a given day, almost 25% are still incubating 10 days later, and more than 5% are still incubating after 20 days. So we consider the converse problem. We consider the data on incubation

periods, and determine a 95% "confidence interval" of time spent incubating. Say that, for example, 95% of individuals have become infective after 10 days. Then we find θ, the mean of the exponential distribution, by solving for θ the equation $\int_0^{10} e^{-s/\theta}/\theta \; ds = 0.95$, giving $\theta \simeq 3.34$. Note that this has the undesired consequence of a slight speed up of processes.

Estimating β is probably one of the hardest tasks in epidemiological modelling, and the value we use is deduced from running simulations repeatedly and observing realistic spread times. In some simulations, we have also used β_i as a parameter to identify. Also useful in determining β is (2): early on in an epidemic, a lot of work is conducted to estimate the value of $\mathcal{R}_0$ using various methods. Using this value, the values estimated for the rates of movement and epidemiological parameters, one can estimate values of β from the expression for $\mathcal{R}_0$ deduced from (2).

Numerical simulations are then carried out using high performance computing resources (HPC) owned by the project as well as HPC resources of Compute Canada. A large number of independent simulations are performed and a number of characteristics of these simulations are computed: number of realizations where the disease becomes extinct, number of realizations where a given city is "hit," i.e., imports an infected case, number of realizations with successful invasion, i.e., where an imported case infects a local individual, etc. Sample results are shown in Figure 3.

7 Conclusions Here, we presented the basic methodological components of the Bio.Diaspora Project. Many extensions of this framework were considered or are under consideration. The model presented here is one where the population is unbounded. Because the time to the next event in the stochastic simulation is exponentially distributed, an unbounded situation quickly leads to a decrease of the time step to an unreasonably small size since the weight of events becomes increasingly large. The first method used to circumvent this problem is to use the so-called τ-leap method [**4**], which allows to consider "packets of events." The newer implementations of our stochastic models, though, use finite populations: the simplifications explained here are not used and the initial susceptible population in the catchment area of airports is the one found using the Dirichlet tessellation method. Standard incidence is used, so that the behaviour of means does not simplify as it does here. Current work in the Bio.Diaspora Project focuses on linking the frame-

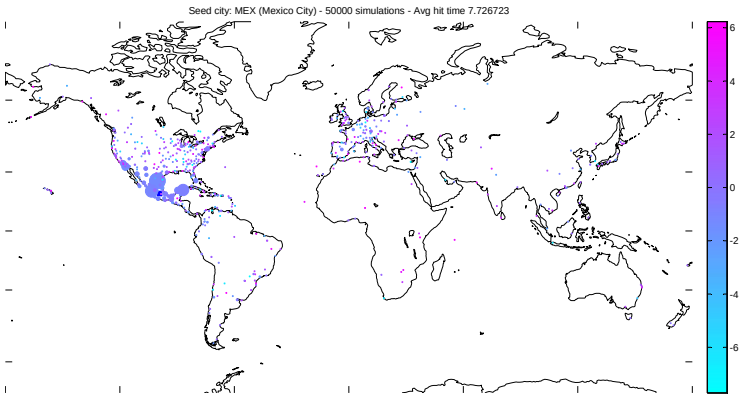

FIGURE 3: Sample result of the numerical simulations. Here, 50,000 realizations were run with a seed in Mexico City (MEX). With the parameters used, the average time when cities were hit conditional on being hit was 7.7 days after the start of simulations. The colour of a dot indicates the average hit time for a given city relative to the overall average hit time, while the size of the dot indicates the number of hits.

work with real-time epidemic surveillance, in particular in the context of a co-occurrence of a major epidemic event and a mass-gathering [**9**].

Acknowledgements Part of this work was initiated during the Fields MITACS Industrial Problem-solving Workshop, held at the Fields Institute in August 2006. Besides JA and KK, individuals involved during the workshop were N. Baddour, C. Breward, A. Gumel, X. Jiang, C. Podder, B. Pourziaei, O. Sharomi, N. Vaidya, J. F. Williams and J. Wu.

REFERENCES

1. L. J. S. Allen, *An Introduction to Stochastic Processes with Applications to Biology*, second edition, Chapman and Hall/CRC, 2010.
2. J. Arino, *Diseases in Metapopulations,* in Modeling and Dynamics of Infectious Diseases, Vol. 11, Series in Contemporary Applied Mathematics, World Scientific Publishing, 2009, 65–123.
3. P. F. Ash and E. D. Bolker, *Generalized Dirichlet tessellations*, Geometriae Dedicata **20** (1986), 209–243.

4. Y. Cao, D. Gillespie and L. Petzold, *Efficient step size selection for the tau-leaping simulation method*, J. Chem. Phys. **124** (2006), 044109.
5. D. Gillespie, *Exact stochastic simulation of coupled chemical reactions*, J. Chem. Phys. **81** (1977), 2340–2360.
6. W. O. Kermack and A. G. McKendrick, *A contribution to the mathematical theory of epidemics*, Proc. Roy. Soc. London, Ser. A **115** (1927), 700–721.
7. K. Khan *et al.*, *An analysis of Canada's vulnerability to emerging infectious disease threats via the global airline transportation network*, Technical report, The Bio.Diaspora Project (St Michael's Hospital, Toronto, Ontario, Canada), 2009.
8. K. Khan *et al.*, *Spread of a novel influenza A (H1N1) virus via global airline transportation*, New England J. Med. **361**(2) (2009), 212–214.
9. K. Khan *et al.*, *Preparing for infectious disease threats at mass gatherings: the case of the Vancouver 2010 Olympic Winter Games*, CMAJ **182**(6) (2010), 579–583.
10. K. Khan *et al.*, *Global public health implications of a mass gathering in Mecca, Saudi Arabia during the midst of an influenza pandemic*, J. Travel Medicine **17**(2) (2010), 75–81.
11. K. Khan *et al. Infectious disease surveillance and modelling across geographic frontiers and scientific specialties*, The Lancet Infectious Diseases (2012).

Department of Mathematics, University of Manitoba,
Winnipeg, MB, Canada.
E-mail address: Julien_Arino@umanitoba.ca

Centre for Research on Inner City Health,
St Michael's Hospital, Toronto, ON, Canada.

Centre for Research on Inner City Health,
St Michael's Hospital, Toronto, ON, Canada.

Centre for Research on Inner City Health,
St Michael's Hospital, Toronto, ON, Canada.

Departamento de Matematicas, E T S I Industriales,
Universidad Politecnica de Madrid, Madrid, Spain.