Hindawi Publishing Corporation
Epidemiology Research International
Volume 2012, Article ID 716072, 18 pages
doi:10.1155/2012/716072

Research Article

Extracting Data from Disparate Sources for Agent-Based Disease Spread Models

M. Laskowski,[1] B. C. P. Demianyk,[1] J. Benavides,[1] M. R. Friesen,[1] R. D. McLeod,[1] S. N. Mukhi,[2] and M. Crowley[3]

[1] *Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, MB, Canada R3T 2N2*
[2] *Office of the Chief Engineer, Canadian Network for Public Health Intelligence, Winnipeg, MB, Canada R3E 3R2*
[3] *Technology Development, MTS Allstream, Winnipeg, MB, Canada R3C 3V6*

Correspondence should be addressed to R. D. McLeod, mcleod@ee.umanitoba.ca

Received 20 October 2011; Revised 8 February 2012; Accepted 5 March 2012

Academic Editor: Jacek A. Kopec

This paper presents a review and evaluation of real data sources relative to their role and applicability in an agent-based model (ABM) simulating respiratory infection spread a large geographic area. The ABM is a spatial-temporal model inclusive of behavior and interaction patterns between individual agents. The agent behaviours in the model (movements and interactions) are fed by census/demographic data, integrated with real data from a telecommunication service provider (cellular records), traffic survey data, as well as person-person contact data obtained via a custom 3G smartphone application that logs Bluetooth connectivity between devices. Each source provides data of varying type and granularity, thereby enhancing the robustness of the model. The work demonstrates opportunities in data mining and fusion and the role of data in calibrating and validating ABMs. The data become real-world inputs into susceptible-exposed-infected-recovered (SEIR) disease spread models and their variants, thereby building credible and nonintrusive models to qualitatively model public health interventions at the population level.

1. Introduction

Complex networks underlie the transmission dynamics of many epidemiological models of disease spread, in particular agent-based models (ABMs). Network-based epidemiological models use a percolation-like principle to simulate disease spread through the population [1], and there are a large number of studies on ABMs and network-based epidemiological models. Agent-based models are of increasing interest due to their potential to capture complex emergent behaviours during the course of a simulated epidemic, where these behaviours arise from the nonlinearities of human-human contacts [2]. ABMs may employ an explicit or implicit social contact network defined by structured agent interactions. In the explicit case, a disease model (e.g., susceptible-exposed-infected-recovered or SEIR type) can be implemented directly on the network. In the case of ABM, these resemble simulation models rather than the steady-state analysis of network-based models mentioned in [1].

In all cases, though, the fidelity of the agent-based framework (model) relies in part on the credibility of the social contact network data that feeds it, defining agents' characteristics, behaviours, and interactions within the model. Potential data sources to define agents include census and demographic data (coarse) and finer-grained data made available by various means of polling personal electronics such as cell phones. In related work it was demonstrated that data to model a social contact network can be collected through web services or wireless sensor devices or "motes" worn by individuals in the target population and subsequently used in an infectious disease spread model [3]. Such an approach has been previously undertaken to gather data, for example, in an organization (workplace or school). The resulting estimated social contact network was used to model an influenza-like illness (ILI) within the setting [4], based on a standard SEIR individual type model. In this time-stepped model, infection spreads between two vertices (individuals) along the weighted edges of the network which represent the degree (frequency and duration, weighted) of social contact between the two individuals. However, estimating fine-grained social contact networks in larger populations (metropolitan scale or larger) through real data sources is an area of research still in its relative infancy and is the interest which is motivating this current work.

In cases where precise contact network data is unavailable, an alternative is to mine data as done by EpiSims [5] which uses United States Department of Transportation information to estimate the schedules of the agents in several metropolitan areas. This presumes that the choices of locations at which agents interact are constrained by the transportation network (model), which itself is a complex network. In EpiSims, schedules for the agents are synthesized from census and USDOT data. A simulation is then run during which a synthetic contact network is constructed from the interactions of the agents and their locations. The resulting dynamic bipartite graph [5] is used to simulate disease spread in the manner stated earlier, except on a much larger scale. Both EpiSims and another well-vetted infectious disease simulator, BioWar [6], initially perform validation on model components separately. This is an important component of plausibly reasoned argument, supporting the statement that the model as a whole functions as specified.

The overall objective of this work is to advance the development of a flexible, accurate, and scalable ABM framework by which to simulate respiratory infection spread within communities of all sizes. This paper focuses specifically on a systematic exploration of five real data sources that have the potential to be integrated into infection spread models (ABMs) to define model topographies as well as agent profiles, behaviours, and interactions to a high degree of accuracy. In particular, the work explores the potential of real data sources as a calibration of the ABM relative to agent behaviour under normal (i.e., nonepidemic) circumstances, as both input into and check of results of an ABM, and as part of the process of building ABMs from the ground up. Thus, this paper situates the exploration of data sources within an infection spread ABM in which the contagions of interest are influenza-like illnesses (ILIs) or other respiratory infections that are primarily contracted through direct or proximal contact. The long-term goal is to develop an agent-based model of high spatial resolution when required, as well as spanning large geographic regions. A further objective is to validate the evolving ABM framework in varying stages of development, by comparing extracted contact networks generated by the ABM to known theoretical or experimental social contact network models.

While the focus of this paper is on real data sources for an ABM, a discussion of the ABM itself is also required. The model developed to date and outlined here is denoted Simstitution, representing the intent to simulate an institution or community through a hierarchy eventually encompassing a province- or state-wide simulation. In this work, the province of interest is Manitoba, Canada covering 649,950 square kilometres. The capital city of Winnipeg, located approximately 100 km from the southern border of the province, is home to 700,000 of the province's 1.1 M residents. Likewise, most of the other 400,000 residents of Manitoba are also located within 200 km of the southern border of the province. Modeling on this scale is complicated by the fact that an ABM needs to account for the state of each agent as well as their interactions. Although extensive effort is made to exploit available data sources, there are still considerable assumptions embedded in the current work. In some cases, the data are used as input into the ABM while in other cases, real data are analyzed, characterized, and compared to similar data extracted from the ABM as a means of validation.

This paper is organized as follows. Five sources of real data are described, in terms of their applicability as inputs into an ABM framework. All data sources are oriented toward the characterization of agents and their behaviours (movements and interactions) during periods of normal (nonepidemic) function and thus serve to calibrate the ABM. These data include (1) fine-grained data from a 76-vehicle travel study; (2) coarser municipal travel survey data; (3) a smartphone app developed within our group, denoted face2face, used to collect personal social network data; (4) cellular communication service records. Subsequently, the ABM framework being developed within our group, denoted Simstitution, is outlined, along with sample simulations. The observed outcomes, especially validation to data sources, and implications are then discussed.

2. Materials and Methods

Materials required for these types of models essentially come down to two entities. The first are data sources and the second is the ABM framework itself, here denoted Simstitution. The data sources can act as inputs into the Simstitution framework, as well as comparative data to validate the output of an ABM simulation.

2.1. Data Sources. It is becoming widely recognized that data being generated greatly exceeds our abilities to process it. This is recognized anecdotally as drowning in a "data tsunami." At the same time, the rise of a "data culture" also affords new and significant opportunities including for microsimulation and ABMs, as opposed to more monolithic analytical tools. Data is becoming available for ABMs oriented to infection spread, including demographics, agent proximities, and agent movement patterns that can all be used to define agent profiles within an ABM at both coarse and fine scales. This section outlines sources that vary in fidelity, each offering their own unique challenges in processing and inferencing. These data sources are representative rather than exhaustive. They demonstrate the implicit value of data originally derived for completely separate purposes, where its applicability to epidemiological modeling is a secondary but invaluable contribution made possible by the increasing availability of data.

2.1.1. Traffic Data Sources

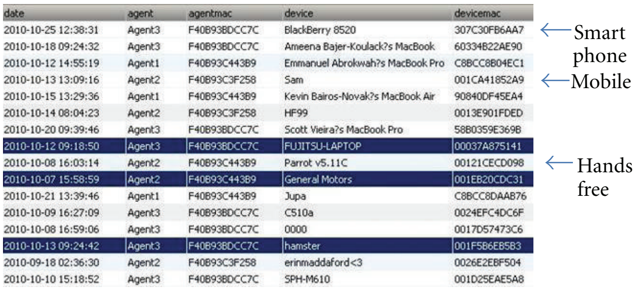
76-Person Probe Vehicle Data. Our group is currently working with two sources of traffic data available for modeling.

One data set consists of approximately 76 individuals whose driving patterns were recorded in extremely high resolution in both space and time for approximately one year. The total number of data points is approximately 44 million and was extracted from [7]. This data in its raw form was anonymized and dereferenced from the city of interest to being located at the North Pole. This data, and similar sources, was primarily used to determine typical driving cycles; however, it also finds application in ABM research. With a few simple heuristics, the data can be easily converted back to its original GPS coordinates. A difficulty with this data—although reasonably voluminous—is that it represents only 0.006 percent of the population and that of a self-selected proportion of the population. However, it does cover a significant amount of time and an opportunity to estimate a person's schedule percentage of time a person may spend at work, leisure and recreation, and so on as well as providing some insight as to variations in circadian and weekly behaviours. These can help guide the agent profiles and behaviours in the Simstitution ABM, as it is essential for any data-driven ABM to have input of this temporal nature.

Municipal Travel Survey Data. The second source of traffic data in our group is derived from a traffic survey made available from the City of Winnipeg Transportation Division [8]. This data and others like it are collected periodically, primarily as an aid in transportation planning. Once available, the data has numerous other applications, including tertiary roles in disease spread models. In this data set, approximately 33,000 users were involved, resulting in over 88,000 trips in private vehicles. The resolution of the available data is the GPS coordinates of nearest intersection for the start and stop timestamps of a trip, and more importantly, labeled with the purpose of the trip. This data has been aggregated within our group and by others into traffic districts which are reasonably close to census districts and, in some cases, very close approximations to census districts. Like the first set of travel data, these data also allow estimates of agent movement patterns into and out of regions. A deficiency of the available data is that it is limited to Winnipeg and surrounding municipalities as opposed to being provincewide. While the data lacks spatial resolution, it carries the benefit of capturing 4.4 percent of the population. Traffic surveys represent an excellent data source and because many are commissioned by public entities (e.g., municipal governments), they are usually readily available in electronic format amenable to mining and analysis.

2.1.2. Cell Phone Data Sources

Face2face Smartphone Application. Our group is working with two distinct sources of data related to personal electronic communications, each with various degrees of fidelity and volume. The first is a smartphone application developed within our group, denoted face2face which is an application developed for Blackberry and Android smartphones that are Bluetooth enabled. The face2face application is designed to poll its local environment on regular intervals for other close-proximity Bluetooth-enabled devices and then record the date, time, and MAC of the discovered device. The application is representative of automated and nonintrusive proximity data collection methods where it is tacitly assumed that consumer electronics serve as proxies for their users. (In this context, "non-intrusive" means that the application requires no active user interaction.) This assumption has limitations, including the disproportionate distribution of cellular devices within a given population to certain demographic subsets; yet, arguably these techniques have increasing credibility as smartphones, and other Bluetooth-enabled personal electronics become more ubiquitous. At the end of 2011, cellphone use was reported to be at 72% of the population, with smartphone sales surpassing feature phones for the first time [9].



date	agent	agentmac	device	devicemac
2010-10-25 12:38:31	Agent3	F40B93BDCC7C	BlackBerry 8520	307C30FB6AA7
2010-10-18 09:24:32	Agent3	F40B93BDCC7C	Ameena Bajer-Koulack?s MacBook	60334B22AE90
2010-10-12 14:55:19	Agent1	F40B93C443B9	Emmanuel Abrokwah?s MacBook Pro	C8BCC8B04EC1
2010-10-13 13:09:16	Agent2	F40B93C3F258	Sam	001CA41852A9
2010-10-15 13:29:36	Agent1	F40B93C443B9	Kevin Bairos-Novak?s MacBook Air	90840DF45EA4
2010-10-14 08:04:23	Agent2	F40B93C3F258	HF99	0013E901FDED
2010-10-20 09:39:46	Agent3	F40B93BDCC7C	Scott Vieira?s MacBook Pro	58B0359E369B
2010-10-12 09:18:50	Agent3	F40B93BDCC7C	FUJITSU-LAPTOP	00037A875141
2010-10-08 16:03:14	Agent2	F40B93C443B9	Parrot v5.11C	00121CECD098
2010-10-07 15:58:59	Agent2	F40B93C443B9	General Motors	001EB20CDC31
2010-10-21 13:39:46	Agent1	F40B93C443B9	Jupa	C8BCC8DAAB76
2010-10-09 16:27:09	Agent3	F40B93BDCC7C	C510a	0024EFC4DC6F
2010-10-08 16:59:06	Agent3	F40B93BDCC7C	0000	0017D57473C6
2010-10-13 09:24:42	Agent3	F40B93BDCC7C	hamster	001F5B6EB5B3
2010-09-18 02:36:30	Agent2	F40B93C3F258	erinmaddaford<3	0026E2EBF504
2010-10-10 15:18:52	Agent3	F40B93BDCC7C	SPH-M610	001D25EAE5A8

Figure 1: A sample of data collected via face2face smartphone application.

A proof-of-concept pilot test has been undertaken with four individuals collecting data on close-proximity Bluetooth-enabled BlackBerry smartphones for just over a three-month period. During this time approximately 500,000 records were collected, where each record is a contact to another close-proximity Bluetooth-enabled device in the general population. Face2face platforms to date include Blackberry Storm and HTC Hero devices (Android). Data includes the MAC and any user assigned metaidentity/type of both the probe device (one of four in the pilot study) and the polled (probed) device, the timestamp, and a location if the probe device is GPS enabled or assisted.

Figure 1 illustrates samples of the data collected and residing on the backend database. Some records provide more information than others, and, as such, several records are perhaps more interesting than others. The second highlighted row indicates a device called General Motors, scanned while the Agent 2 probe was on a local highway. Many other devices are much more easily identified and more easily associated with actual persons. Culling of Bluetooth devices that are not obviously a person is possible but has not been undertaken here at this time. The number of records that were not phones or personal mobile appliances was small in comparison to the total number of records.

In addition to the MAC address and personal metadata, Bluetooth also provides a "class of device/service" (data not shown here) application programming interface (API) which can be further used to differentiate the agent scans. For example, a "class of device/service" entry may take the form of 0x40020C, which a publicly available Bluetooth standard identifies as a telephony service and a Smartphone device.

The face2face contact data is conjectured to be a type of data that can be described by empirical laws. The distribution used follows the Pareto law. Pareto's law is given in terms of the cumulative distribution function (CDF); that is, in this case the number of contacts (N_c) with duration larger than or equal to a duration is an inverse power of the duration as expressed below:

$$P[N_c > D] \sim D^{-p}. \tag{1}$$

From the Pareto distribution, a power law exponent was calculated and varied from 1.4 to 1.75 for the four probe devices used (R^2 values were consistently above 0.95). A power law exponent less than 2 implies that there is no first moment or mean associated with the distribution. As the data obtained from the probe devices is finite, a mean can be calculated, though.

An interesting but not surprising parameter that can be extracted from the Pareto principle is the 80/20 rule. From the data collected, the 80/20 rule was applied to indicate the number of contacts that comprised 80% of the total contact duration. From this, it was estimated that 80% of a person's time is spent with a number of personal contacts that varied between 7 and 20, for the four probe devices. This was extracted from the number and duration of contacts with approximately 5,000 unique Bluetooth devices probed. This is consistent with intuition that although the total number of daily contacts may be large, the majority of one's time is spent with only a small number of people. Within the ABM, this information is useful as it represents a parameter that can be used to generate a person's cohort group within an institution or workplace. The Pareto distribution is directly related to Zipf's law which is more easily calculated for this type of data. An example of a simple ABM using this type of data for contact distributions is illustrated in sequel.

Cellular Service Data. The second source of cellular data analyzed consisted of data provided by a cellular service provider (MTS Allstream). The data consists of all network requests over four days, including the cell tower GPS and antenna sector (if applicable) that the mobile device is associated with, the AAA record (every time the phone accesses the network excluding voice and SMS), and timestamp of the access. These data typically provide service providers with input for network planning, investments, and management of evolving needs. These data also have considerable application to public health interests, although at this time it is difficult to derive its direct benefit in contrast to more explicit inputs such as those associated with census and demographic data, due to both technology and policy issues.

In this work, four consecutive weekdays in November 2010 were extracted from the MTS Allstream dataset. Even at four days, this represented just over 14 GB of data. These data were processed, and an hourly record of anonymized user trajectories was generated. The number of users was approximately 182,000, representing approximately 15 percent of the population. The trade-off in these data is that that spatial resolution is at the scale of an antenna sector. Typically, however, antenna sectors tend to correspond to community areas where census demographic data is also readily available from Statistics Canada. The data used here represent just one of several telecom service providers in the province of Manitoba, some of which share MTS' towers as well as operating their own. For epidemiological modelling, it would be desirable for providers to share anonymized subscriber trajectory data with epidemiological modelers. In terms of contact-based infection spread modeling where movement of (and intersection between) individuals in place and time is of paramount importance, cell phone trajectories are likely the best source of data on a large scale and will remain so for the foreseeable future.

2.1.3. Census Data: Statistics Canada Data Sources. Data sources available from Statistics Canada are the most obvious and only mentioned here for completeness. An unprecedented amount of detail has accumulated from census participation [10]. Within Winnipeg, details are associated with urban neighborhood clusters, and these clusters are further refined into neighborhoods with considerable levels of detail related to households, dwellings, modes of transportation, and so on. This is absolutely essential information for agent characterization in an ABM or microsimulation, as the agents need to have the phenotype of actual persons. Correspondences between traffic districts, neighborhood clusters, and cell tower sectors are not isomorphic but they are similar, which is fortuitous when combining these data sets. These province-wide community profiles also exist for Manitoba Rural Municipalities, Cities, Towns, Villages, Large Government Districts, Indian Settlements, and Indian Reserves in Manitoba. The latter is very important for modeling of respiratory infection spread, as census information provides information related to overcrowding and conditions that contribute to an outbreak. It is also recognized that First Nations people have an incidence of underlying chronic medical conditions that is higher than the national average, putting them at increased risk of severe illness from respiratory infection [11].

2.2. The Agent-Based Model. While the focus of this paper is on the potential and applicability of four real data sources, the data are contextualized within an infection spread ABM, and thus the discussion includes the ABM itself as the second element in ABM microsimulation. The model described in this paper is a project milestone in the process of designing and implementing an ABM simulation framework geared towards high-fidelity modeling of human institutions of varying scales. The framework, inclusive of the four data sources, is new; individual pieces of the framework and/or data sources at earlier stages of development have been used in previous publications. The ABM framework, denoted Simstitution, has broad design goals based on the collective experience of the authors while developing context-specific agent-based models of human institutions. Originally, models of hospital emergency departments [12] and cities [13] were implemented upon "one-shot" simulators, that is, a simulator strongly coupled to the specific modeling application [14]. A one-shot simulator is comparatively easy to implement and gives the modeler fine control over the simulator processes, enabling them to fulfill their requirements. Typically, in order to minimize development effort, the designer will make assumptions which ease the implementation of the model at hand, without consideration for how these assumptions will constrain or complicate repurposing the simulator to implement a different model. From a software engineering perspective, part of the reason that one-shot models are so easy to produce is that little or no effort goes into making the software reusable or extendible. The large number of one-shot simulators observed in the literature [14] is problematic because by their nature they are difficult to reuse. The reusability of the simulator in turn affects the reliability of the simulator. The more researchers that (re)use a particular simulator, the more chances that bugs will be identified and fixed. Furthermore, when a number of models produce reasonable results using a common simulator, confidence in the simulator is increased. Publishing results from a series of models built upon a common simulator framework, combined with verification of model components (or submodels), is a common path for building confidence in simulator frameworks for epidemiological modeling [5, 6].

2.2.1. Simstitution Design Goals. Although there are several frameworks [15–21] which can be used to develop agent based models, these are dwarfed by the number of one-shot or otherwise domain-specific simulators, suggesting that no framework has yet hit upon a "sweet-spot" between flexibility, specificity, extendibility or scalability, and specific support classes for human-centric domains [14]. Human-centrism includes the notion that agents are spatially oriented and situated, since humans are physical entities that occupy and traverse space, rather than existing in some abstract information domain. Simulator support for a range of human time steps on the order of seconds to hours or days is also desirable. We have used commercial simulators such as AnyLogic [19] for building ABMs. In this case, a custom simulator was used as it offered advantages in terms of exploiting hierarchy, data fusion, and parallelization (grid or cloud computing).

Other design features include adherence to software engineering principles to improve reuse and maintainability of the framework, as well as extendibility especially where machine learning can be leveraged for automated generation of agent policy [22].

For rapid model construction, a next generation of ABM framework should facilitate the incorporation of real-time data such as from a database, leading to increasingly data-driven simulation. A tool for visualization and interacting with the model in a graphical manner (GUI) also facilitates model development, validation, and debugging. Visualization is also key for communicating results with subject matter experts and stakeholders [23]. Such a visualization tool can also be extended to serve as a tool for model construction or editing model parameters imported from real data.

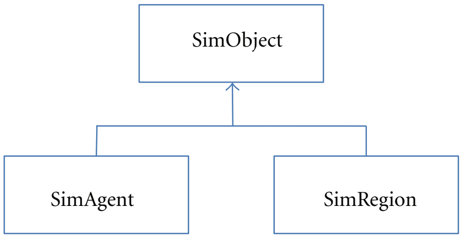


Figure 2: Class diagram for core Simstitution class hierarchy.

The accessibility of agent behavior development to persons with a nonprogramming background can be improved by first providing a scripting layer on top of the compiled code and then perhaps adding a visual or block (e.g., OpenBlocks [24]) programming (drag and drop) on top of that. Over time, a library of useful scripted behaviors can be built up.

The increasing availability of parallel or distributed computing systems also suggests that contemporary or future agent-based simulator frameworks should support distributed, parallel, or cluster computing. The increasing availability of cluster-based computing resources (a consequence of Moore's Law), sensitivity to real-time computational constraints, and medical data privacy issues augur well for cluster-based computing. As a result, the Simstitution framework design emphasizes scalability with respect to multiple processors and discrete memory spaces over efficiency in executing one particular type of model.

Naturally limiting the degree of accessibility of the environment limits what agents can perceive and interact within the environment (including other agents). Localizing agent perception not only fits in well with the agent paradigm but also limits to what extent information needs to be shared between processes in a distributed model, which should facilitate using spatial decomposition as a guide for distributing computational load.

These disparate goals will require balance in feature choice and design.

2.2.2. Simstitution Design Details. Simulated entities within Simstitution fall into either of two major categories: agents (SimAgent), which are the autonomous entities that make decisions and interact with the environment, and instances of the SimRegion class, which represent spatially partitioned subdivisions of the environment. Note from Figure 2 that the SimObject is abstract and exists because SimAgent and SimRegion have much of their interfaces in common.

One of the core design tenets of Simstitution is that the spatial division is closely intertwined with the division of computational work across processors and discrete memory boundaries. Therefore, SimRegion is unit of spatial decomposition as well as a convenient unit of computation. In the latter role, it can be considered as a container for agents that need to have their next state computed. Figure 3 illustrates the details of this relationship. A particular instance of SimRegion can be the parent container of SimAgents or

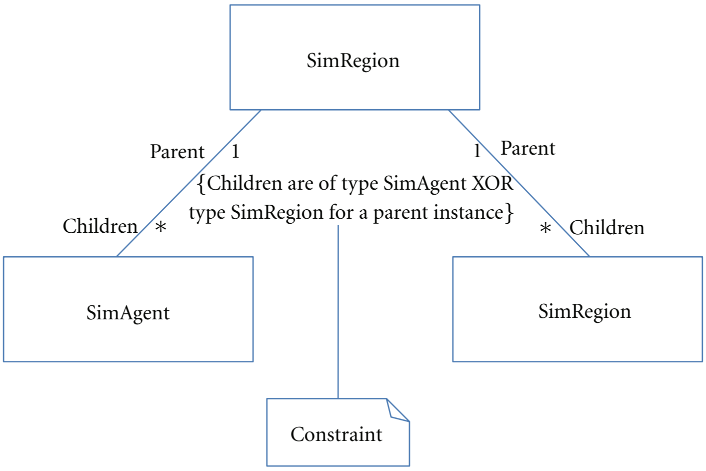

SimRegion
Parent 1
1 Parent
{Children are of type SimAgent XOR type SimRegion for a parent instance}
Children *
* Children
SimAgent
SimRegion
Constraint

Figure 3: Relationships between core class instances, forming a tree.

SimRegions but not both types at the same time. This restriction will in practice result in tree hierarchies of SimRegions, with SimAgents contained in the leaf SimRegions and the "top region" at the root of the tree. The SimRegion spatial decomposition granularity becomes increasingly fine as it moves away from the root and towards the "leaf regions" of the tree.

Time advances in the simulation when the simulator advances the time of the top region (root of the tree) by some discrete time step. The top region will then advance the time of its children by the same time step in a recursive fashion such that the tree is traversed in a depth-first manner, until all the SimAgents in the leaf regions have been simulated for that time step. The simulator will restart this process again, until a certain number of time steps have elapsed.

IndividualPolicy is a modular unit that affects the behavior of the subscribed SimAgent, which may also require the IndividualPolicy to store encapsulated SimAgent state data specific to that IndividualPolicy. Examples are a schedule policy which causes the SimAgent to observe a particular day/night work/home schedule or, in the case of a hospital being modeled, a doctor policy which causes the SimAgent to treat patients within a hospital. Within a SimRegion, each possible concrete-derived IndividualPolicy class has a corresponding GroupPolicy for that SimRegion. The GroupPolicy acts as a factory for the corresponding IndividualPolicy and, if required, facilitates coordination between one or more derived IndividualPolicy classes (e.g., healthcare worker policy in a hospital that coordinates interaction between nurse and doctor IndividualPolicies). Implicit here is the assumption that the properties of the local environment constrain the behavior of agents (e.g., airport security lineup, swimming pool, hospital, bank, etc.). The associations between SimRegion, SimAgent, GroupPolicy, and IndividualPolicy are shown in Figure 4.

Communication or interaction between SimAgents exclusively uses messages passed between SimAgents. Messages received by a SimAgent are relayed to its IndividualPolicies which can lead to an internal change of state or an action to be taken which could lead to additional messages being sent to other IndividualPolicies on the same subscribed SimAgent or messages sent to other SimAgents. Message passing fits well with the agent paradigm, since the alternative implies a direct mapping between external events and internal agent state which violates the principle of agent autonomy [25]. It is in the message passing and the IndividualPolicies that changes in individual agent behaviour are triggered in response to agent-specific characteristics combined with specific external conditions. This models infection control and mitigation strategies by modeling individual behaviour change or behaviour management during infection outbreaks.

3. Results and Discussion

This section provides visualizations of the data sources and how they can be applied within the ABM to model agent behaviours relative to movement patterns and interactions. The following discussion then demonstrates the ABM using aspects of the data in microsimulation scenarios.

3.1. Visualizations: Traffic Data

3.1.1. 76-Person Probe Vehicle Data. Within the detailed 76-person (probe vehicle) traffic data itself and from a summary report of these data, one is able to discern areas of interest, such as parking lots frequented. These are typically associated with airports, malls, places of work, and residences. The value of these data is the extraction of duty cycles for various routine activities. This level of detail is required with an ABM, as each agent is basically operating on a schedule which may be potentially interrupted depending upon their health state within each agent's SEIR (susceptible-exposed-infected-recovering) stochastic infection model.

An example of a concatenation of two trips is illustrated in Figure 5, demonstrating the geomapping of the data. Reverse engineering of one trip indicates that the destination is a fitness centre and the second trip originates from

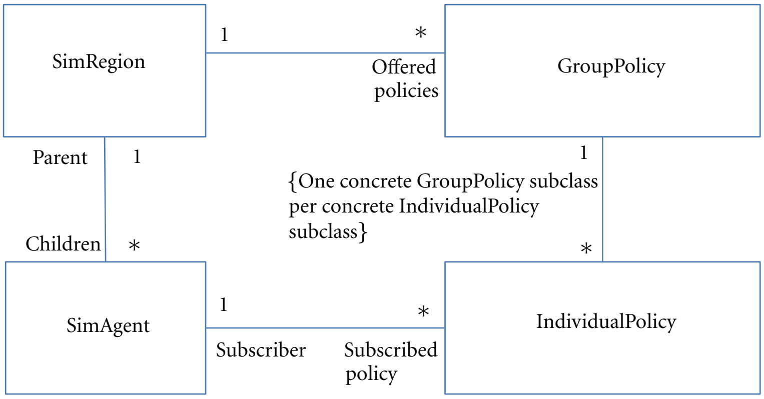


Figure 4: Relationships involving modular agent policies.

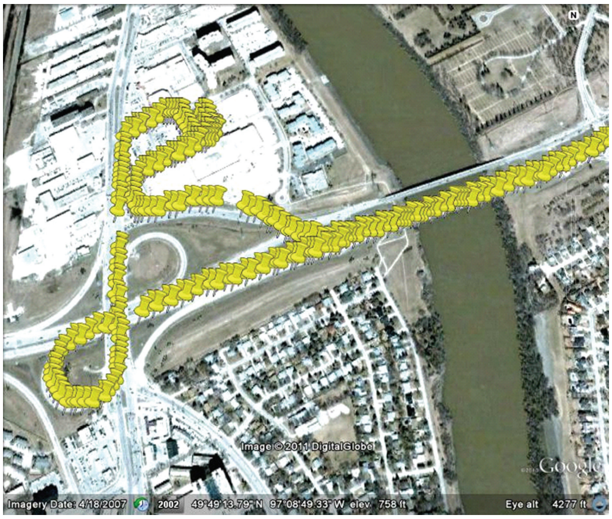


Image © 2011 DigitalGlobe
Google
Imagery Date: 4/18/2007
2002
49°49'13.79" N 97°08'49.33" W elev 758 ft
Eye alt 4277 ft

Figure 5: Geomapping of probe vehicle same data: a person visiting a local commercial facility.

the fitness centre. Automating the inference of activity is error prone, but an estimate of activity is sufficient for the ABM as we are not concerned with any specific individual, but rather a prototypical individual. While the origins and destinations may be inferenced from the data, the identity of the person or persons involved remains anonymous and unknown to the researchers.

From this type of data, an estimate of the most common parking lots and time durations spent there can be compiled. Parking lots serve as proxies for the institutions (e.g., workplaces, schools, leisure facilities) with which they are associated. To some degree, these data have already been similarly analyzed by others, as one of its original intents was to use the data to estimate where PHEV recharging stations could be ideally located. For our ABM purposes, the interest is in extracting duty cycles associated with an agent's schedule of activity in order to add credibility to assumptions made within the model.

3.1.2. Municipal Travel Survey Data. From the data available within the Winnipeg Area Travel Survey (WATS) (33,000 users), we are able to generate coarser flow maps within the main urban center (Winnipeg) and its surroundings. Again for visualization purposes, a limited number of trajectories are overlayed on Google maps using GPSVisualizer [26] and illustrated in Figure 6.

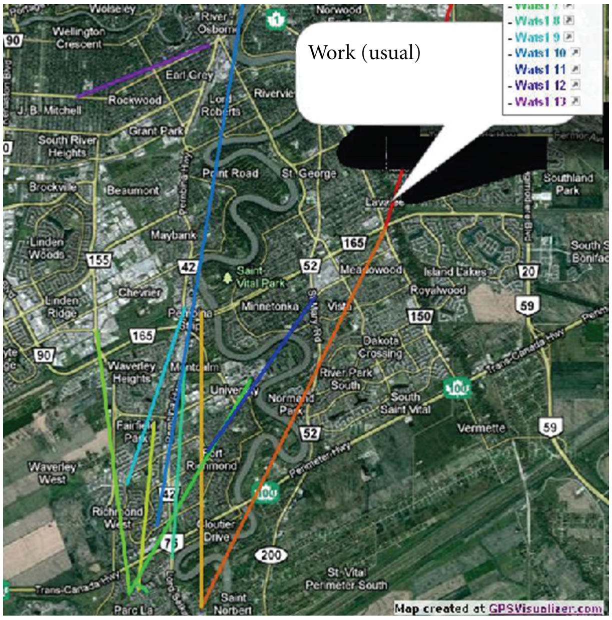


Figure 6: Geomapping of travel survey sample data: origin and destination.

Although considerably coarser than data from probe vehicles, the data in the WATS dataset is labeled in such a way that one is able to determine the purpose of trip, mode of transportation, and number of persons in the vehicle. As the number of trips is approximately 88,000, over time this provides a reasonable estimate of intracommunity flow from a macroscopic perspective. This data is relatively easy to generalize and allows for reasonable models of flow, peak and nonpeak times. The value of the data relative to the ABM is to use the data both as an input and as an instrumented output. The flows within the ABM should ideally resemble those extracted from the WATS dataset.

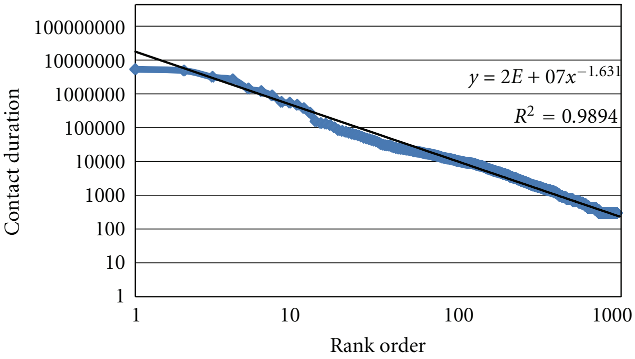


Figure 7: All agents (aggregated) ranked in order of contact duration.

3.2. Visualizations: Smartphone Data

3.2.1. Face2face Smartphone Application. From the face2face smartphone application, one can analyze data from at least two perspectives. The first is a measure of rank ordering of a person's contacts. In this manner, a Pareto distribution and/or directly related Zipf's law exponent can be extracted. The results of this analysis are summarized as follows.

Figure 7 illustrates the rank ordering aggregated over all agents in the pilot study of the face2face app running on four probe devices. The rank order exponent (Zipf's law) is approximately 1.63. This yields an estimated power law exponent of approximately 1.61. The implication is that an agent's contact pattern would follow a power law distribution (heavy tail) without finite moments. This result is expected from both the face2face application pilot study as well as intuitive perceptions of real face-to-face contact patterns. The value of this type of measure is for credibility checks after-ABM simulation. As contacts can be easily instrumented within the ABM, similar contact pattern distributions are to be expected.

A second application of the face2face application is simply logging the GPS of cellular-assisted location services. An example is shown in Figure 8. The left hand side is simply the uploading of GPS coordinates as a person traverses a local golf course. Samples were taken every 30 seconds and are sufficiently fine for most applications. The right hand side includes a visualization of contacts as well as GPS data. The contacts are merely represented as being in concentric circles and do not have any further spatial meaning. This type of data—although easily collected with Smartphones—would still require participant buy-in. The face2face application using Bluetooth tracking has been described in considerably more detail elsewhere [27]. Furthermore, we are in the process of extending the face2face application in a manner that would explicitly facilitate and encourage use and sharing of the data.

3.2.2. Cellular Service Data. Cellular data made available from MTS Allstream was analyzed to explore its use in infection spread modeling. The data consist of timestamps of anonymized users and associated cell tower sectors. The data was processed to provide a given cellular phone's most likely associated cell sector tower, once per hour. This parameter was input into the ABM and in this manner can provide a coarse movement pattern (trajectory) of the person for whom the cellular phone is a proxy. Example trajectories are illustrated in Figure 9, each representing approximately four days. For visualization purposes 330 trajectories are illustrated.

Table 1: Self-entropy estimates.

Time period	Average Entropy	Towers Visited
00:00:00 to 07:00:00	0.773	1.71
07:00:00 to 18:00:00	1.770	3.41
18:00:00 to 23:59:59	1.393	2.63

In Figure 9, the majority of activity is within the main urban centre, with approximately 10% of the trajectories in rural and northern regions of the province of Manitoba. The number of trajectories is approximately 182,000, with approximately 1/3 of those being more or less complete (when a phone is off or a person is roaming, there is no local record of the associated cell tower, and the agent trajectory is interrupted). This type of data can be directly used within the ABM, although still only representative of a sample of agents with a degree of inherent self-selection bias. Analysis of the data provides meaningful measures in order to parameterize schedules for prototypical agents within an ABM simulation. Some preliminary analysis was to treat the trajectories as antenna sector generators and measure their entropy for various times of day. Figure 10 illustrates the entropy for users between 00:00:00 h to 07:00:00 h and from 07:00:00 h to 18:00:00 h. The entropy is calculated as the self-entropy and is a measure of the number of towers a user would typically visit during a given period of time. Higher entropy correlates with greater mobility over a greater geographic area. Self-entropy is defined here as

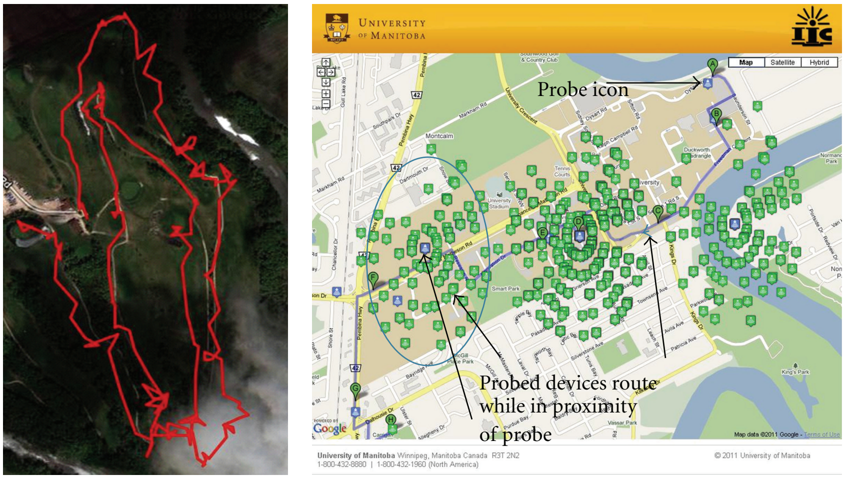
$$\text{Self} - \text{enropy} = -\sum_{\text{celltowers}} p_i \lg(p_i), \qquad (2)$$

where p_i is probability that a cell phone user is associated cell tower/sector i, providing a measure of the number of towers a user visits and where lg is $\log_2$ providing the measurement in bits.

For Figure 10, 12,000 trajectories were sampled at random, and the self-entropy overnight and during weekday business hours was generated. The 7:00 to 18:00 h time period captures both early morning and afternoon peak traffic periods.

An interpretation of the self-entropy measure is as follows as shown in Table 1.

As expected, the activity measured in terms of cell sector towers visited during the day is greater than the activity in the evening, with the least activity occurring overnight. In all three time periods, there is a pronounced peak at entropy of zero, that is, one cell tower repeatedly recorded for a given cell phone, indicating that the person (cell phone) remained within that specific cell tower sector. The aggregate


University of Manitoba
Map
Satellite
Hybrid
Probe icon
Probed devices route while in proximity of probe
Google
Map data ©2011 Google
University of Manitoba Winnipeg, Manitoba Canada R3T 2N2
1-800-432-8880 | 1-800-432-1960 (North America)
© 2011 University of Manitoba

(a)

(b)

Figure 8: Sample face2face data: agent trajectory (a) and close-proximity discovered contacts (b).

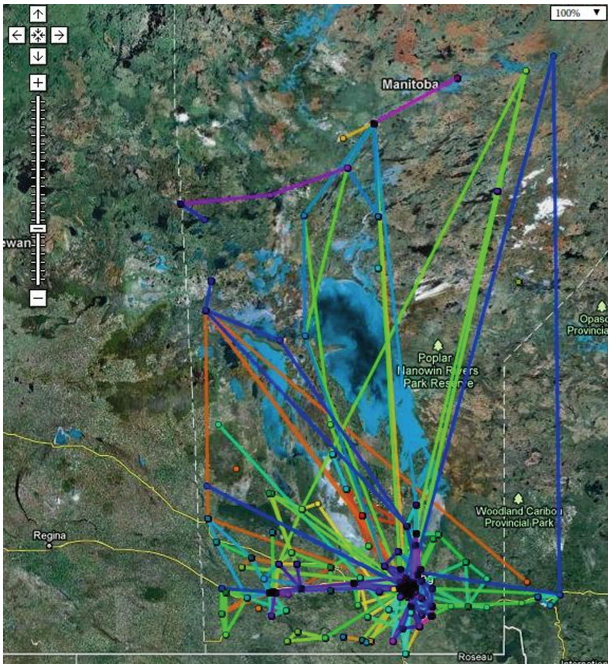


Figure 9: Geomapping of cellular record sample data: 300 cellular phone trajectories by cell tower.

entropy was also calculated in the manner that included all cell tower records from all users. For 07:00:00 h to 18:00:00 h, the aggregate entropy was 6.99. The implication is that the majority of the records were for approximately 128 sectors out of just over 450 possible sectors. During the night, fewer than two towers were recorded while during the peak period, 3.41 towers were visited. Apart from using the data directly to represent specific agent schedules, this can also be instrumented *from* the ABM and can again be used as a check on model credibility.

3.3. Geographical Integration: Communities, Cell Towers, and Traffic Districts. While each data set uses slightly different descriptions for regional areas, they follow similar patterns roughly guided by population. Figure 11 illustrates the alignment of geographical regions as parameterized from the different types of data available.

A decision was made to standardize on cell tower sector areas, as they span most of the province and reasonably closely capture similar areas as the census and traffic sectors. One challenge is that the density of tower sectors in some regions (e.g., downtown Winnipeg) greatly exceeds a neighborhood or a traffic district, and some interpolation is required.

3.3.1. Summary of Data Sets. Table 2 provides a comparison of the data sets in terms of their coverage, spatial resolution, potential for real-time acquisition, and their potential for enhancement through technologies like Bluetooth.

In comparing data, the most promising is that associated with cell phone trajectories, and in the future an expectation would be that the % covered would increase and the associated spatial resolution decrease. However even in the present case of low spatial resolution the cell phone trajectory data tracks data from the municipal travel survey data set (in a comparison of person trips in and out of the core versus time of day) [28].

We note that while the size of some of respective data sets is sufficient to infer statistical representation of the respective populations, such a generalization is nonetheless premature since not all details of the respective sampling strategies are known. In the case of the 76-person probe vehicle data set, the sample was very small and further biased by the use of self-selected volunteers [7]. The municipal travel survey (WATS) data are reasonably voluminous and were derived from a telephone survey with associated biases and constraints of availability and willingness to participate [8].

Table 2: Data comparisons.

Data set	% of Population	Spatial resolution	Real-time potential	Bluetooth potential
Probe vehicles	0.006 (Province) 0.01 (city)	Very high (m)	Somewhat	Yes
Municipal travel survey	2.75 (Province) 4.4 (city)	Medium (100 m)	No	No
Face2face smartphone application	Negligible for pilot study; potential is ubiquitous	Very high (m)	Yes	Yes
Cell phone trajectories	15	Low (km)	Yes	Yes

Table 3: SEIR modeL health state transition diagram parameters estimates.

Parameter	Value
Δ_E	Duration of incubation period (e.g., 24 hours avg)
Δ_{IW}	Duration of infectious period (at work), for example, 2–5 days avg
Δ_{IH}	Duration of infectious period (at home), for example, 3 days avg
Δ_R	Duration of recovered (immune) period, for example, 200 days avg
α, p	Contact graph transmission probability from home contacts
β, p	Contact graph transmission probability from work contacts

The face2face Smartphone application data are the least representative but hold considerable promise for scalability and illustrate how fine grained contact resolution is technologically feasible. The cellular service data cover a substantial fraction of the population but also carry inherent bias. While the cellular service data used in this work were anonymized, market studies [29] provide typical demographic profiles of Smartphone users that could be incorporated into subsequent ABM simulations.

3.4. ABM Simulations. Simulations were carried out to explore the potential of the data sets as constituent components of an ABM and as a means of calibrating normal agent movement patterns in normal, nonepidemic periods.

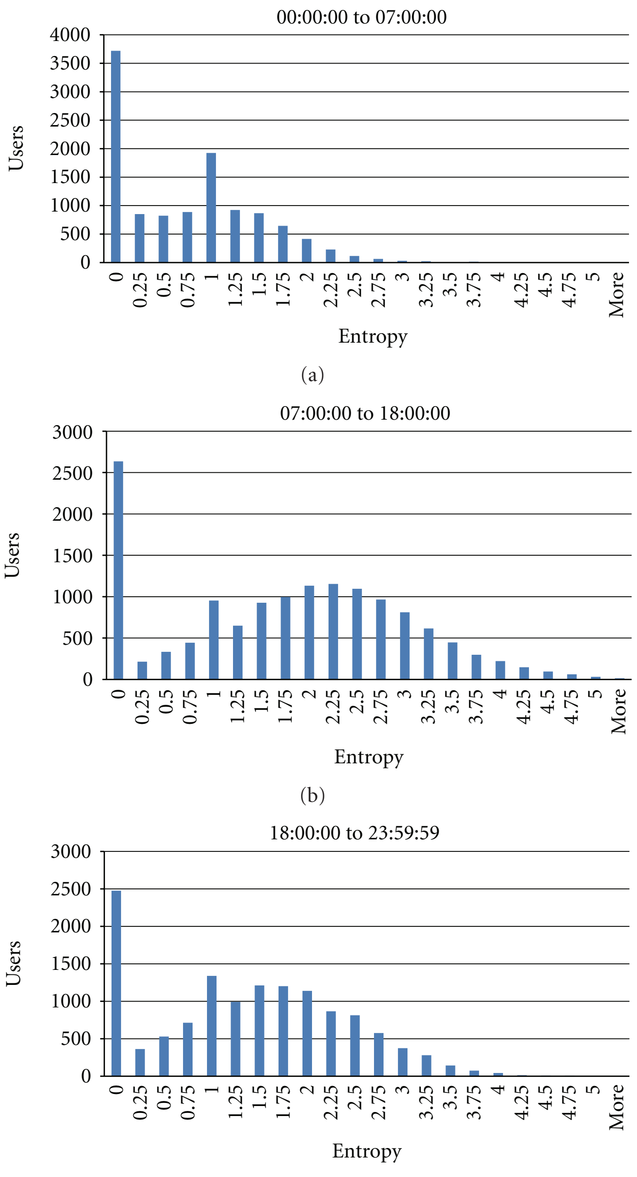
3.4.1. Simulation 1: 5000-Agent, Isolated Community. Initial ABM simulations within the Simstitution framework began with a consideration of integrating the face2face application data. The model developed is, in essence, an individual-based model. Using notions of the 80/20 rule extracted from the data collected, an SEIR infection spread model was built and run, simulating the dynamics of an influenza type illness (ILI) similar to pH1N1 of 2009/2010. In the case of Manitoba, isolated northern communities were particularly hard hit during the first wave of the pH1N1 outbreak, and, hence, a community in relative isolation is of interest for simulation. Although the work is not attempting to replicate any specific community, the population considered was a model with 5000 people in relative isolation. This also provides a closed system for modeling purposes.

The model used as a base was a simple SEIR agent-based or discrete model. It is a phase-type model where an individual can be in one of a number of health states, typically denoted susceptible, exposed, infected, and recovered. This is a minimal type phase space and is illustrated in Figure 12. Parameters associated with the model are shown in Table 3.

In this work, the infected state consists of two phases (work and home). In general, a person may be infected and infectious at work prior to a period where they may be ill and at home (immobile with probability q). Each person has essentially two contact lists: one associated with their day-to-day business activities with parameters governed by the 80/20 rule and their home contact list consisting of more direct family members. Using data collected from the four face2face application probe devices, each person (agent) had a contact list of approximately 10 close contacts, reflecting the 80/20 rule found from the Pareto distribution associated with contacts. The simulation—although coarse—includes a circadian rhythm where each individual was also provided with a contact list of a small number of persons during the night (every second 12 h cycle), in addition to their daytime contacts. The probability of becoming infected was $P = 0.005$. This was implemented as a 0.005 probability of becoming infected if one of your close contacts was infected, per hour of contact with that agent. This measure of infection probability is an adjustable parameter within the simulation but is consistent with considerably larger models.

Figure 13 illustrates the spread of an infection through a population of 5000 persons, with curves typical of compartmental susceptible-infected-recovered (SIR) models. The only significant difference here is that these simulations are the result of individual stochastic models with contact lists governed by the observation of the 80/20 rule arising from the Pareto distribution of inferred contacts from an automated proximity contact pattern generator. In epidemiology, Ro is denoted the basic reproduction number of the infection and is the number of secondary infections a single infected case will cause. In the case of an influenza strain (e.g., 1918) Ro has been estimated to be between 2 and 3. In the data of Figure 13, Ro is approximately 1.9.

To further explore conditions that may be representative of remote northern communities, which also tend to have lower average personal income levels, the number of close proximity contacts during the "at home" cycle was varied from 2 to 5, as illustrated in Figure 14. This represents a tendency towards overcrowding in homes. Qualitatively, the simulations indicate that overcrowding is a major contributing



(a)

(b)

(c)

Figure 10: Self-Entropy of users for various times of day.

factor in the spread of an ILI. The overcrowding exacerbates the infection spread as a consequence of increased exposure due to increased agent-agent contact [30]. This finding makes one common public health intervention inappropriate in that context—that being the recommendation for an infected individual to stay home. In environments with severe overcrowding in homes, this recommendation may in fact be deleterious. In these scenarios, it may be worth investigating the impact on overall community infection rates, if temporary mobile facilities are established to house and treat infected individuals.

3.4.2. Simulation 2: "Morden Simulation" of 16,500 Agents in Adjacent Communities. Towards continued evolution of the ABM under the Simstitution framework, the next iteration saw the development of a model approximating a small community, inclusive of census data and cellular service records, as well as a degree of spatial behavioral modeling. Toward the development of a provincewide simulation, our work developed a small-scale ABM of two adjacent communities in the Rural Municipality of Stanley, Manitoba with a combined population of approximately 16,500 residents: Winkler, Manitoba at 10,000 residents and Morden, Manitoba at 6500 residents. This is a spatial temporal model with demographic data from Statistics Canada [10]. From this perspective, agents are provided with schedules, and a model of disease spread is run. Figure 15 illustrates the topography of the region of interest.

The towns of Morden and Winker are roughly seven miles apart in southwest Manitoba. One of the reasons for selecting this geographical area is that it is representative of many North American rural municipalities. Figure 15 also illustrates the location of three cellular service towers with MTS Allstream as the service provider. The ABM is discussed in terms of model validation using data that is mined from anonymized cell phone use records. In addition to cell phone usage, the model is also improved using the face2face Smartphone application to provide high-fidelity contact patterns.

This simulation incorporated the ABM framework features outlined earlier, as well as visualization capabilities to observe emergent model behavior during execution. In this model, the SimRegion tree only consists of two layers: the root or top SimRegion (Morden) and the leaf SimRegions which represent the home, school, and work locations that agents occupy. The leaf SimRegions are arranged in a grid with empty spaces between structures to allow for SimAgent travel. Agents are assigned work, school, and home locations based on demographic data [10].

Figure 16 shows a screenshot of the Morden simulation at a particular time step. On the left side, the entire city is shown. On the right side is a detailed view of six classrooms in a school in the center of town in which individual SimAgent details can be seen. Details include the gender and age of the SimAgent, as well as disease status. Disease status is the most interesting and is indicated by the color of the SimAgent icon. The icon changes color, with green indicating a susceptible state. Once the agent is infected, it turns yellow, orange, and red depending on how long they have spent in the infected state. Finally, recovered SimAgents turn blue. The leaf SimRegions are depicted as colored squares. SimRegions with no SimAgents contained inside are white. SimRegions with one or more SimAgents display a blended color tile based on the aggregated disease state of the SimAgents inside. For example, the green tiles (Figure 16) indicate that most of the agents within that tile are in a susceptible state. Further, the purple tile is due to the large number of recovered agents present at that location. This type of visualization is very useful to allow a user to watch a time sequence of the simulation and observe movement and infection spread. In effect, this becomes part of the debugging process.

Four concrete IndividualPolicy subclasses were used to generate the SimAgent behavior in the Morden model. The SchedulePolicy determines whether a particular agent wants to be at its assigned work, school, or home, depending on

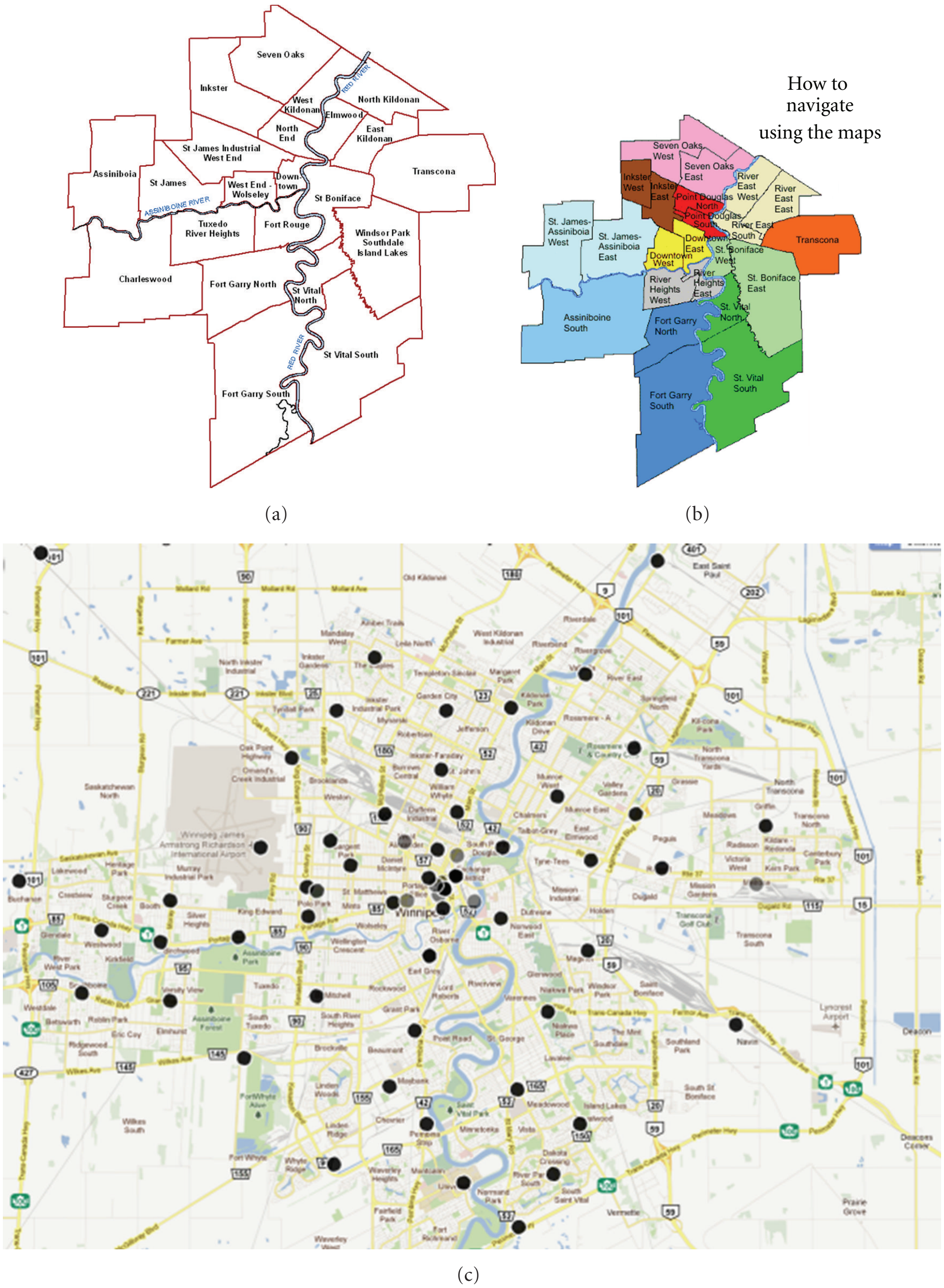


Figure 11: Traffic districts (top left), census regions (top right), and dell tower locations (bottom) for Winnipeg, Canada.

the demographic profile of the particular SimAgent and the current time which advances in increments of one hour. The SchedulePolicy sends messages containing the desired destination to the SimAgent's MovementPolicy which handles the actual movement. The InfluenzaPolicy maintains the particular SimAgent's disease state and, if in the infected state, sends "infection" messages to other SimAgents in the same SimRegion, which is how disease spreads between SimAgents. Finally, the BluetoothTrackingPolicy emulates the face2face Smartphone app and is the source of the synthetic contact data. Currently, the corresponding GroupPolicies were used to facilitate aggregation of data in a spatially explicit manner to achieve the tiling effect in Figure 16.

Thus, the data sources in this simulation include demographic data, coarse grained data from anonymized cellular records, and the finer grained face2face Smartphone application programmed to log close-proximity Bluetooth device contacts. In this simulation, the face2face Smartphone app

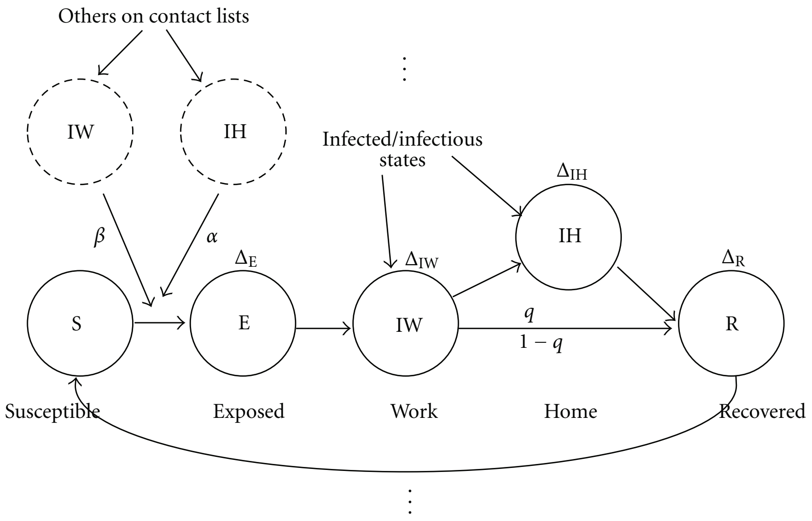


Figure 12: The stochastic process governing an individual's health state in the SEIR model.

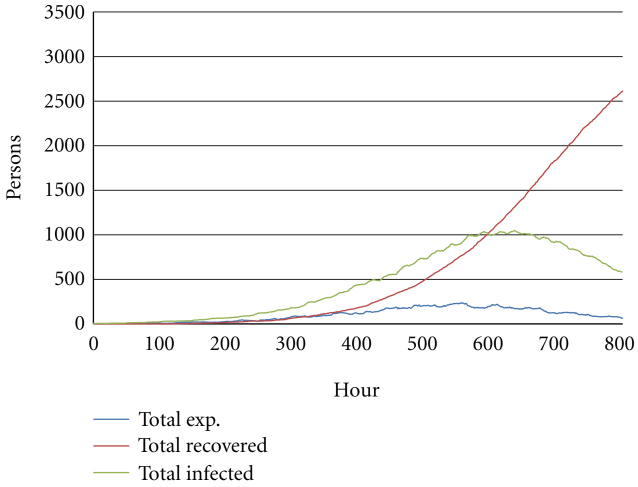


Figure 13: Infection spread simulation using contact data extrapolated from Table 3 $P = 0.005$ and population of 5000.

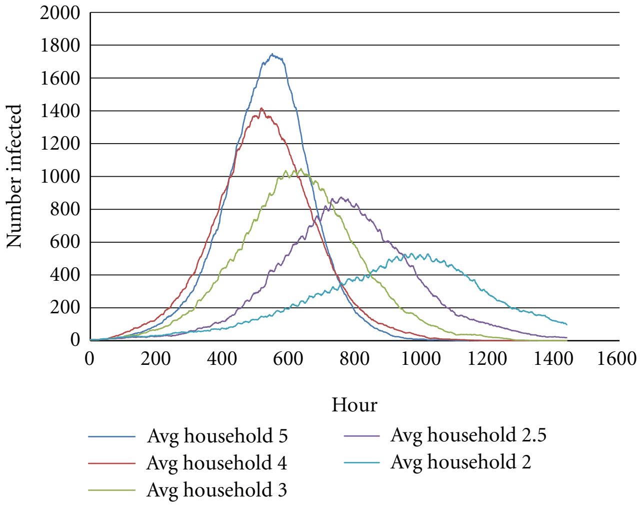


Figure 14: Infection spread simulation assessing the impact of overcrowding.

data was used as a verification check. In essence, the ABM was instrumented with tracking capabilities and contact distributions generated for comparison with those measured.

Once processed for the connections with the towers of interest (Figure 15), this amounted to just under 500,000 records of the overall four-day dataset from MTS Allstream. Although statistical in nature, the data can be further processed to estimate flux of persons between the two neighboring towns. Within an infection spread model, this helps in estimating patterns of movement that contribute to infection spread. Once stored in a database, queries allowed for extracting anonymized device activities. Figure 17 illustrates the breakdown of mobile cellular devices accessing the cell towers in Morden and/or Winkler. For an individual, a duty cycle can be estimated, illustrating the percentage of time a person is likely to be in one region or another. The timestamp can also be used to infer primary community of residence. User counts here indicate that approximately 2650 users remained in Morden, approximately 485 users remained in Winkler, while 2285 users spent time in both Morden as well as Winkler over the four-day data collection period.

This data can be refined further based upon those with cellular network access records in both Morden and Winkler. Figure 18 illustrates the breakdown of users who accessed cell towers in both communities over the duration of a single connection of their cellular device to the network, implying travel from one community to the other community during the time period of the device's network connection. The actual device accesses between the two communities break down as approximately 65/35, reflecting durations more accurately.

3.5. Validating and Evolving the ABM Based on the Outcomes of the Morden Simulation. In addition to their role as inputs

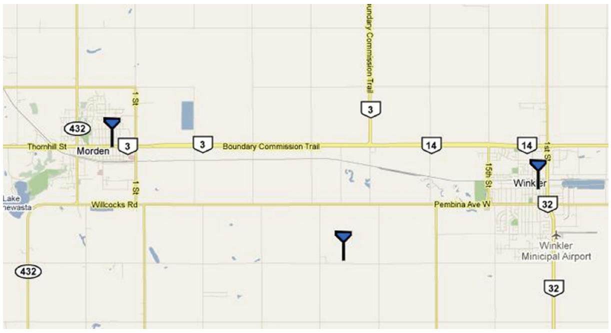


Figure 15: ABM model topography of adjacent small communities: Winkler, Manitoba and Morden, Manitoba.

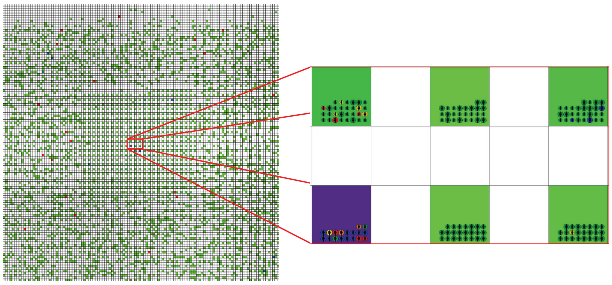


Figure 16: Screenshot of running simulation. Morden (left), close-up of 6 classrooms (right).

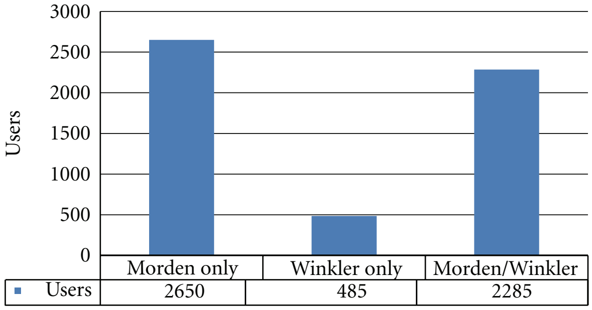


Figure 17: Morden and/or Winkler cellular user aggregates.

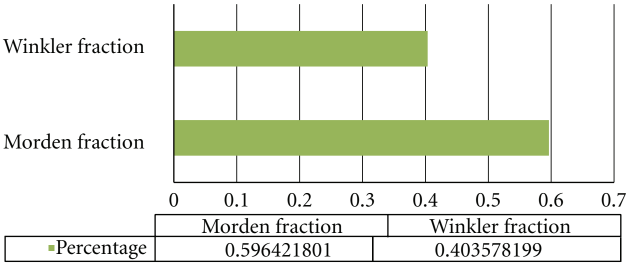


Figure 18: Breakdown of users with cellular records in both communities.

Table 4: Zipf exponents for various demographics.

Age	Zipf exponent	R^2
2	1.85	0.85
6	1.91	0.82
12	2.23	0.81
16	1.89	0.87
20	2.28	0.86
30	2.26	0.91
40	2.16	0.91
50	1.97	0.91
70	1.35	0.90

governing and calibrating agent behaviour in ABMs, real data sets can also validate models, in this case the Simstitution ABM framework. The first and most obvious would be using as accurate demographic data as possible. The Morden simulation relied on data obtained through the federal census by Statistics Canada, and the Simstitution framework in general is designed for the inclusion of these federal census data. In addition, models of schools in the Morden simulation were refined to provide for reasonable class sizes, data which are estimated here but would benefit from using real data. With this model, a disease spread simulation was run and provided a baseline for modeling the spread of a respiratory infection or ILI. Figure 19 illustrates the infection spread within the Morden simulation.

In the first effort to improve the basic ABM within the Morden simulation, it was instrumented in terms of agent contacts and durations which should be validated by reflecting the patterns in data extracted from the Bluetooth probe devices. The objective was to see how well the model reflected real person-person networks. For the baseline simulations in the Morden simulation, contact patterns for all agents were instrumented. From these simulations and the aggregated rank orderings, an 80/20 rule can also be estimated. In this case, 80% of the contact durations are spent with approximately 4% of a person's contacts (25/670). This again is consistent with data extracted from the face2face Smartphone pilot study. Figure 20 illustrates the rank ordering of contact parameterized by demographics. Intuitively these profiles appear reasonable. School age children spend considerable time with three groups, household members, school classmates, and friends. The knee in the curve of school age children is between 20 and 32. For samples of age groups, the exponents associated with Zipf's law are presented in Table 4. Perhaps it is also intuitive that a 2-year old and a 70-year-old person have similar—and somewhat limited—contact patterns.

The consequence of the rank ordering implies that the coefficient associated with the corresponding Pareto distribution would be between 0 and 1. The lack of a finite mean in the corresponding contact PDF approximation would imply that a few long duration contacts are a significant vector of infection spread.

Other means of validating the data from a simulation includes its relation to other types of published data. For example, in [31] contact patterns are analyzed as derived from a large population survey that indicated that for their preliminary modeling "5- to 19-year-olds are expected to

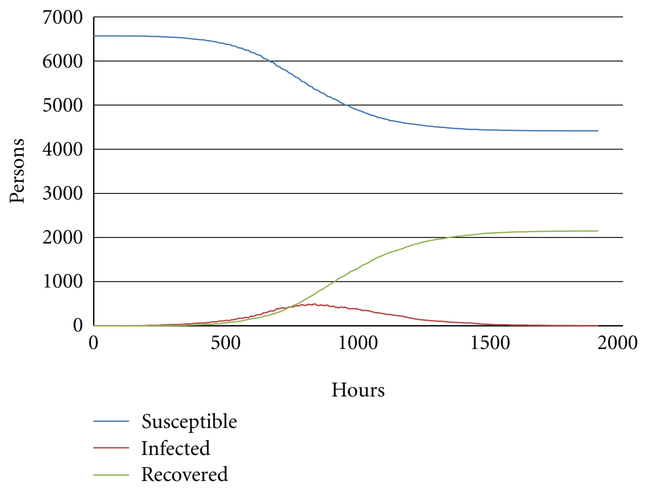


FIGURE 19: Susceptible-infected-recovered (SIR) disease spread simulation.

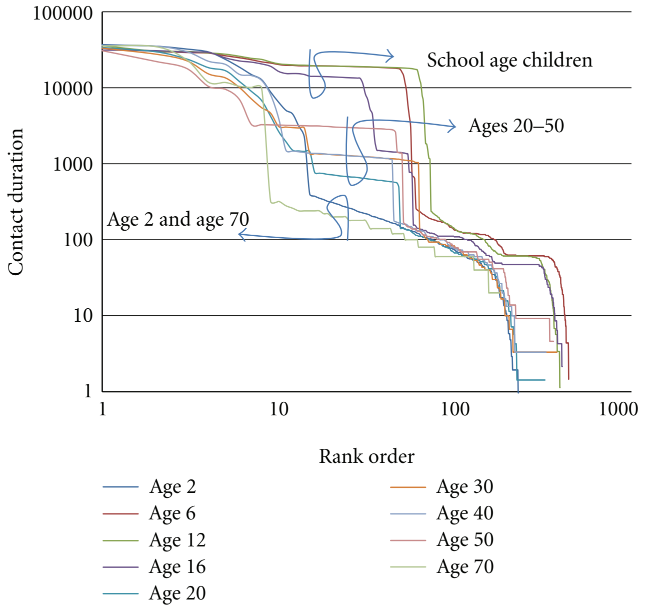


FIGURE 20: All agents ranked in order of contact duration.

suffer the highest incidence during the initial epidemic phase of an emerging infection transmitted through social contacts measured here when the population is completely susceptible.” These expectations are consistent with the contact patterns generated by our ABM.

In the second instance of enhancing the ABM based on the observed outcomes of the Morden simulation, it was recognized that Morden does not exist in isolation, and, as such, flux of persons into and out of the area is required. This is not unlike large-scale efforts where simulations are based upon data extracted from airline travel, for example. In this case the data—albeit voluminous—is reasonably extractable. It is more difficult to obtain inter-community travel in rural settings as one cell tower may cover the entire town. In this environment, there are few (if any) directly available data sets; however, there are opportunities for inferencing from more disparate data sources. Although an ABM running a bounded topography may be applicable to geographically isolated communities, in semirural settings there is considerable interaction with surrounding towns that need to be accounted for. From Figures 17 and 18, an indication of interactions between Morden and Winkler can potentially be inferred from cellular tower access. The data suggests that of the cell-phone-carrying persons (approximately 4000) with primary residence in Morden, approximately 34% are seen to have records in both Winkler and Morden, with that person spending on average 65% of their time in Morden and 35% in Winkler. Similarly of the approximately 1400 phone-carrying persons with primary residence in Winker, approximately 65% are seen to have records in both Winkler and Morden, with that person spending on average 65% of their time in Winkler and 35% in Morden. These very coarse estimates nonetheless allow one to begin modeling multiple communities and their interactions.

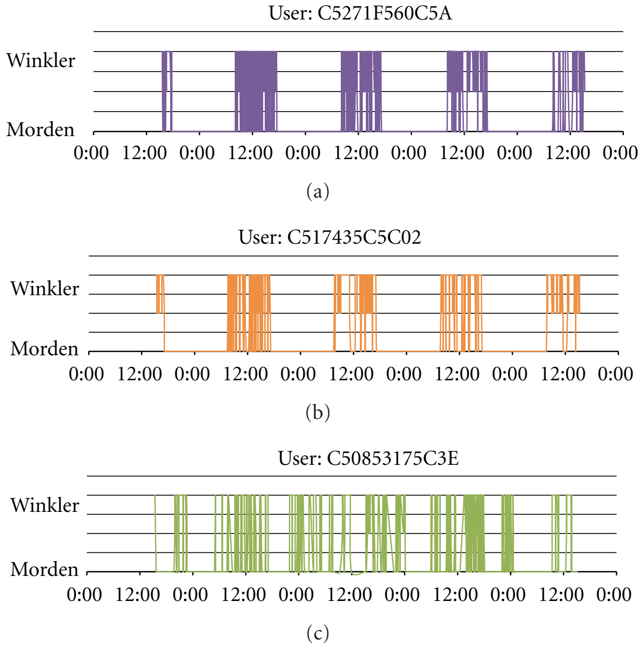


FIGURE 21: Temporal sequence diagram of three users accessing cellular towers in Morden and Winkler.

One can burrow deeper into the data and determine periods of time a representative individual would spend in each community. Figure 21 illustrates a typical daily duty cycle associated with randomly selected users and their access to cellular towers in Morden and Winkler. The first two user data duty cycle plots reinforce routine activity theory as users are primarily seen in Morden during the night with intertown tower records primarily during the day. The third user’s behavior is considerably more erratic. In either case, these types of trajectories are required in improving interacting ABMs and improving the specificity and accuracy of agent behaviours and interactions within an ABM framework.

Ongoing ABM framework evolution is depicted in Figure 22 where external sources are integrated as they become available. At present, these are done in a manual fashion but are amenable to automation and/or machine learning, and thereby further adapting the model to the real world.

A benefit to developing an ABM framework in this fashion is that it provides opportunities for increasing levels of computational efficiencies by exploiting parallel computing paradigms, since many communities tend to be autonomous with limited interaction between communities. The next significant milestone in the development of a provincewide simulation using the Simstitution ABM framework will be to decompose major urban areas, utilizing the municipal travel survey data for gross flows and generating trajectories for all agents based on generalized but detailed profiles extracted from the cellular record data. Cohort groups will be correlated to that of the face2face Smartphone app data, and schedules of activities will be guided by detailed trip data.

4. Limitations and Opportunity

There are a number of limitations in attempting to incorporate real data from somewhat disparate sources into an ABM framework, including two primary challenges in fusing data to a model, regardless of data source. The first challenge is the collection of the data, with assurances that the data collected is meaningful and accurate, and then mining or interpreting the data for parameters or characteristics useful to the model. The second challenge relates to integrating the data into the model itself, running simulations, and then attempting to qualify (and ideally quantify) the outputs. In many instances, the results of the simulations may be self-fulfilling and somewhat self-evident, as vercrowding in isolated and impoverished communities leads to increased infection spread. The infection spread interventions that one could model may provide guidance for policies that may then be considered. For example, an intervention associated with reducing infection spread may be a recommendation to stay home while ill; in overcrowded settings, a more effective intervention may be quarantine or a modified quarantine policy whereby an infected person may be advised to seek temporary housing in a facility set up specifically for that purpose. This may also not be unobvious, but modeling with real data may help to elucidate and verify these options and interventions [32].

Ideally one would like to compare the output of a disease spread model with major outbreaks. For a number of reasons this is not always possible. The purposes of models are to aid in understanding how effective the anticipated public health interventions will be in the event of future outbreaks. As such, when using ABMs, an objective is to make the models as accurate as possible using real data to the greatest degree possible. This is one of the major advantages of using ABM, in that they lend themselves to inclusion of real data which is correspondingly becoming increasingly available. Although not modeled here, there is also a significant medical facility intermediate between Morden and Winkler in the Morden simulation, providing an effective vector for infection spread as both patients and health care workers largely come from both Morden and Winkler. These low-level details and data are required as one attempts to generate a comprehensive disease spread ABM on the scale of a province or state.

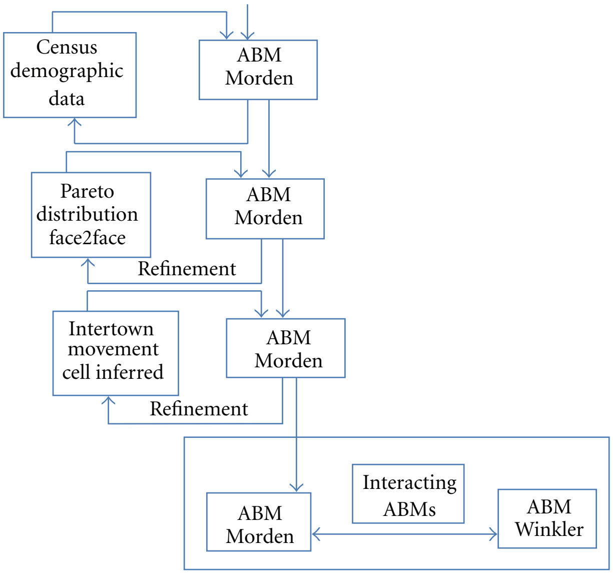


Figure 22: SEIR disease spread simulation.

The opportunities for using real data within an ABM and related microsimulations are considerable. There are obvious challenges in dealing and interpreting data but the access and availability and the availability of computer power to mine and analyze data have never been greater. The ABMs under development in our group are not limited to ILIs but are also amenable to other contact-based infections such as STIs. In that case, modifications would include extensions to the stochastic health state process associated with each individual, as asymptomatic states would need inclusion. It could be that to simulate STI transmission, an ABM would be particularly suitable, as it would have to attempt to account for infection predispositions within various demographic profiles; this is something that would be difficult within a more general modeling method. A related project in our group is to build a hybrid ABM combining behaviors and movements extracted from real data sources such as those described here, in conjunction with the interaction with electronic social networks at particular sites, with the express objective of finding real-life social connections. An advantage of ABM microsimulation in this application would be the ability to apply mock infection interventions and estimate their efficacy.

Work in progress includes developing the simulation of SEIR model of infection spread within the Simstitution ABM framework based on agent contact data extracted from cellphone trajectories in a provincewide simulation scenario (188,000 agent trajectories interpolated to 1.1 M agents). Each antenna sector GPS coordinate represents a location where persons can come into contact with one another, given their cellphone trajectory. Limitations to this model are somewhat self-evident, as the cellular data have biases not yet taken into consideration, as well as the rather obvious fact that even though agents may be within the same antenna sector range, it is unlikely that all agents are in proximate contact. Cohort profiles extracted from the face2face Smartphone application data can be used to generalize the contact profiles. Another challenge within the pending simulation is associated with agent susceptibility, which is a consequence of many factors including economic and environmental conditions such as overcrowding. However, in model building—much like many design problem solving initiatives—a divide-and-conquer paradigm is often a reasonable approach. Our longer term-goals are to acquire real-time trajectory data feeds and integrate these into the Simstitution ABM framework in a way that may also facilitate prediction and simulation of infection mitigation policies.

The value of this type of data fusion within an ABM is closely related to Stein's phenomena, which implies that as more data is added, the estimators tend—on average—to be more accurate.

5. Summary

This work has explored the potential of real yet disparate data sources in an agent-based modeling framework for simulation of infection spread within populations. The data serve as calibration for agent profiles (behaviours and interactions) during normal, nonepidemic periods, and the use and cross-references of multiple data sets improve the credibility and validity of the model. The data sources included a Smartphone application (face2face) that estimated proximate contacts and durations to similar devices, cellular service records that allow one to estimate a person's trajectory, municipal travel survey data, and fine-grained trip data from a pilot study of 76 vehicles over one year. The unique contribution of the work is the integration of technologies that generate real contact data with existing sources of real contact data into the ABM framework, then applied to govern infectious disease spread models.

Acknowledgment

The authors thank Tony Florio, at Research in Motion for an equipment donation of five BlackBerry Storm devices. Financial support from Manitoba Hydro is gratefully acknowledged.

References

[1] M. E. J. Newman, "Spread of epidemic disease on networks," *Physical Review E*, vol. 66, no. 1, Article ID 016128, 2002.

[2] J. M. Epstein, "Modelling to contain pandemics," *Nature*, vol. 460, no. 7256, p. 687, 2009.

[3] B. Demianyk, D. Sandison, B. Libbey et al., "Technologies for generating personal social network contact graphs," in *Proceedings of the 12th IEEE International Conference on e-Health Networking, Application and Services (HealthCom '10)*, Lyon, France, July 2010.

[4] M. Salathé, M. Kazandjieva, J. W. Lee, P. Levis, M. W. Feldman, and J. H. Jones, "A high-resolution human contact network for infectious disease transmission," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 51, pp. 22020–22025, 2010.

[5] P. Stroud, S. Del Valle, S. Sydoriak, J. Riese, and S. Mniszewski, "Spatial dynamics of pandemic influenza in a massive artificial society," *Journal of Artificial Societies and Social Simulation*, vol. 10, no. 4, 2007.

[6] K. M. Carley, D. B. Fridsma, E. Casman et al., "BioWar: Scalable agent-based model of bioattacks," *IEEE Transactions on Systems, Man, and Cybernetics A*, vol. 36, no. 2, pp. 252–265, 2006.

[7] S. Shahidinejad, E. L. Bibeau, and S. Filizadeh, "Winnipeg Duty Cycle: WPG02," 2010, http://mspace.lib.umanitoba.ca/handle/1993/3997.

[8] Winnipeg Area Travel Survey, http://transportation.speakupwinnipeg.com/WATS-Final-Report-July2007.pdf.

[9] 2012, http://www.washingtonpost.com/business/technology/smartphone-shipments-surpass-feature-phones-in-europe/2011/09/08/gIQAKjemCK_story.html.

[10] 2012, http://www.statcan.gc.ca/.

[11] 2010, http://www.sixnations.ca/H1N1FluVirus_Updated%20-Fact%20Sheet290909.pdf.

[12] M. Laskowski, R. D. McLeod, M. R. Friesen, B. W. Podaima, and A. S. Alfa, "Models of emergency departments for reducing patient waiting times," *PLoS ONE*, vol. 4, no. 7, Article ID e6127, 2009.

[13] M. Borkowski, B. W. Podaima, and R. D. McLeod, "Epidemic modeling with discrete space scheduled walkers: possible extensions to HIV/AIDS," *BMC Public Health*, vol. 9, supplement 1, p. S14, 2009.

[14] A. Uhrmacher and D. Weyns, Eds., *Multi-Agent Systems: Simulation and Applications*, CRC Press, New York, NY, USA, 2009.

[15] 2010, http://www.simsesam.de/.

[16] 2010, http://wwwmosi.informatik.uni-rostock.de/mosi/projects/cosa/james-ii/.

[17] S. Luke, C. Cioffi-Revilla, L. Panait, K. Sullivan, and G. Balan, "MASON: a multi-agent simulation environment," *Simulation: Transactions of the society for Modeling and Simulation International*, vol. 82, no. 7, pp. 517–527, 2005.

[18] 2010, http://spades-sim.sourceforge.net/.

[19] 2010, http://www.xjtek.com/.

[20] 2010, http://www.swarm.org/.

[21] J. H. Miller, "Active nonlinear tests (ANTs) of complex simulation models," *Management Science*, vol. 44, no. 6, pp. 820–830, 1998.

[22] M. Laskowski, *An agent based decision support framework for healthcare policy, augemented with stateful genetic programming [Ph.D. thesis]*, University of Manitoba, 2010.

[23] E. Bonabeau, "Agent-based modeling: methods and techniques for simulating human systems," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 3, pp. 7280–7287, 2002, http://www.pnas.org/content/99/suppl.3/7280.full#xref-ref-3-1.

[24] http://education.mit.edu/openblocks.

[25] H. V. D. Parunak, "'Go to the ant': engineering principles from natural multi-agent systems," *Annals of Operations Research*, vol. 75, pp. 69–101, 1997.

[26] GPSVisualizer, http://www.gpsvisualizer.com/map?output_google.

[27] J. Benavides, B. Demianyk, R. D. McLeod, M. R. Friesen, K. Ferens, and S. N. Mukhi, "3G smartphone technologies for generating personal social network contact distributions and graphs," in *Proceedings of the IEEE International Conference on*

Healthcare Informatics, Imaging and Systems Biology (HISB '11), San Jose, Calif, USA, July 2011.

[28] R. Neighbour, M. R. Friesen, R. D. McLeod, S. N. Mukhi, and M. Crowley, "Vehicular traffic modeling governed by cellular phone trajectories," Submitted VTC 2012.

[29] 6/12: Cell Phone Nation, http://maristpoll.marist.edu/612-cell-phone-nation/.

[30] A. J. McMichael, "Environmental and social influences on emerging infectious diseases: past, present and future," *Philosophical Transactions of the Royal Society B*, vol. 359, no. 1447, pp. 1049–1058, 2004.

[31] J. Mossong, N. Hens, M. Jit et al., "Social contacts and mixing patterns relevant to the spread of infectious diseases," *PLoS Medicine*, vol. 5, no. 3, article e74, pp. 0381–0391, 2008.

[32] M. Laskowski, L.C. Mostaço-Guidolin, A.L. Greer, J. Wu, and S. M. Moghadas, "The impact of demographic variables on disease spread: influenza in remote communities," *Scientific Reports*, vol. 1, article no. 105, 2011.