An Open Data Driven Epidemiological Agent-Based Model for Irish Towns

Elizabeth Hunter[1], Brian Mac Namee[2], John Kelleher[1]

[1]School of Computing, Dublin Institute of Technology, Ireland
[2]School of Computer Science, University College Dublin, Ireland
elizabeth.hunter@mydit.ie
brian.macnamee@ucd.ie
john.d.kelleher@dit.ie

Abstract. Agent-based models have become popular as tools for epidemiological simulations due to their ability to model at the individual level. To produce results that can be readily applied to a given population, agent-based models need to be data rich and it can be difficult to obtain sources of data representing the population being modeled. There are, however, many sources of openly available data that can be used to create a simulation. In this paper we demonstrate how openly available data, primarily from Ireland's Central Statistics Office, can be used to develop epidemiological simulation models for Irish towns. Using openly available data allows anyone to have access to the data used in the model to evaluate or recreate our work. We present case studies modeling two different towns, Rathnew and Bagenalstown.

1 Introduction

Agent-based models are a type of computer simulation composed of agents that can interact with each other and with an environment. An agent can be anything from an individual to an organization or body, such as a nation state. Agents will make decisions on what to do and how to interact with other agents based on a behaviour control program. The behaviour control program can be different for each agent, or for different agent types [11]. Many different fields use agent-based models including ecology, demography, geography, political science and epidemiology [12]. Although agent-based models have been around for some time, with one of the earliest published models appearing in 1971, it was not until the late 1990s with the introduction of various platforms designed to help create agent-based models (Netlogo, Swarm and Repast) that they began to gain popularity in the social sciences. The modelling platforms allow for non computer programmers to create and understand agent-based models making them available to more researchers. As the platforms improve and computer power expands more and more models can be created that would have been nearly impossible to create before [11].

Although agent based models are common in epidemiology, the spread of infectious diseases through a population is traditionally modeled using equation

based models. One of the main disadvantages of equation-based models is that they assume every agent in the population has an equal probability of coming into contact with every other agent in the population. This is referred to as homogeneous mixing [10]. Agent-based models, on the other hand, allow what is referred to as heterogeneous mixing as the probability of an agent coming into contact with other agents can vary based on factors such as age, gender, and occupation. This is allowed because agent-based models work at the agent or individual level and so can model interactions and social networks with much more fidelity to the real world. It is this heterogeneous mixing that makes agent-based models so useful for epidemiology [4].

An important part in setting up an agent-based model is collecting the data needed to simulate the population and the society. Although some models use minimal data, for example Dunham [9] describes a simulation where agents are ageless and genderless, the majority of agent-based epidemiological models are data rich. Data rich models produce relevant results that can aide researchers in planning for future outbreaks or learning from past outbreaks. In these types of simulations demographic data is required to simulate the population being modeled. Typically census data is used to build the population [2, 16, 4, 21, 3, 15]. To have a more accurate distribution of households across the environment some models also use population density data [19, 20]. Geographic data such as land use data, elevation data or street maps can also be used in an agent-based model [18, 3, 8, 5, 15, 14].

The amount and types of data used in creating epidemiological agent-based models varies. One of the differences is the data available to the researchers: if cell phone location data is available agent movements will be much more accurate than if the model only uses census transportation data [10]. If the researchers have access to fine level population density data the distribution of agents will be more accurate compared to only using high-level census data [20]. However, many data sets that could be useful in creating a data rich agent-based model are not openly available to all researchers. Being able to create an accurate model from openly available data has the advantage of being adaptable to multiple populations and societies and being able to be recreated and used by anyone who has access to the data.

In this paper we show how publicly available data can be used to create epidemiological agent-based models for Irish towns. The paper outlines the data used, how the data is put together to simulate the population and how the data is used to control how the agents move and interact with each other within the model.

2 Data Sources

We use the computer software Netlogo [24] to implement the simulations described in the paper. Netlogo is an easy to use and popular environment for creating agent-based models [11]. It does, however, have disadvantages, one of the biggest being the speed of the program. When modelling simulations with a

small number of agents Netlogo works well, however, once the number of agents gets large enough the simulation slows down. As we decided to model small Irish towns with populations of about 3,000 Netlogo is sufficient for our use. The open-data-driven approach described in this paper, however, could be used with any agent-based modelling tool or platform.

In order to create the model different types of data are needed including population statistics, GIS data and workplace and school locations. The majority of data used in the model comes from Ireland's Central Statistics Office (CSO) but other sources are also used. The following sections outline the sources of the data used in the model.

2.1 Population Statistics

The CSO provides a wealth of open access data. The data is taken from the results of the Irish census which occurs every five years. The data currently available from the CSO is from the 2011 census. Data can be downloaded at multiple geographic levels and is organized into fifteen different themes each with a set of tables containing information on the population. The themes are described in Table 1.

Theme 1: Sex, Age and Marital Status	Theme 9: Social Class and Socio-Economic Group
Theme 2: Migration, Ethnicity and Religion	Theme 10: Education
Theme 3: Irish language	Theme 11: Commuting
Theme 4: Families	Theme 12: Disability, Carers and General Health
Theme 5: Private Households	Theme 13: Occupation
Theme 6: Housing	Theme 14: Industries
Theme 7: Communal Establishments	Theme 15: PC and Internet Access
Theme 8: Principal Status	

Table 1. The 15 themes from the CSO census data tables [7]

For the purpose of this simulation, the data can be downloaded at the small area level. Small areas are areas of population that contain between 50 to 200 dwellings. They are designed as the lowest level of geography that the CSO uses for compiling statistics [6]. The download will result in a csv file with the data for all small areas in Ireland containing data for each table within each theme. The small areas related to a town being simulated and the necessary tables can be selected from the csv file. The small area boundary file discussed in the next section provides a mapping between small areas and towns. Table 2 contains the information on the different CSO tables that were used to create the simulation.

Theme-Table	Table
1-1	Population by Sex and Age
4-2	Family units with children by size and age of children
4-3	Family units with children by type of family and age of children
5-1	Private households by type
5-2	Private households by size
8-1	Population aged 15 years and over by principal economic status and sex

Table 2. The tables from the CSO data that were used in creating the model [7]

2.2 GIS data

A GIS dataset containing the small area data was downloaded from the CSO website. The CSO provides access to boundary files from the 2011 census. The files contain the boundaries at different levels including provinces, counties, electoral divisions, towns and small areas [6]. The data set downloaded from the CSO website contained small area information for all of Ireland: the QGIS [22] software was used to select only the small areas that overlapped with the town being simulated so the data could be loaded into Netlogo. The small area boundaries do not always match town boundaries, thus the small area dataset could potentially cover more area than the town being simulated.

2.3 Other Sources

While the CSO data provides most of what is necessary to create the model, the data does not provide information on workplaces or schools. The Department of Education and Skills in Ireland has data on individual schools, including enrollment [1]. In the dataset the address of some schools are given while for other schools only the town in which they are located is recorded. Thus in order to determine where to place these schools their locations are determined using Google maps[1]. Locations of any business parks are also found on Google maps and placed in the simulation based on those locations. The number of work places in a town can be found on local business pages website[2]. Although adding schools and workplaces into the model is currently done manually, it may be possible to automate the process by geocoding the locations and using the geocode to place the location in the model. Land use data could also be used to determine the locations of industrial areas where workplaces could be located.

3 Simulation

To create the simulation the data is used to setup the town in the modelling environment, populate the town with agents and create a schedule of movement for the different agents. This section describes what is involved in each of those steps and how an infectious disease model can be added to the simulation.

3.1 Town Setup

The town was setup using the small area datasets. The number of occupied households in each small area can be determined using household tables from the CSO. This number is then used to randomly generate the correct number of household locations within each small area. Using a combination of Google maps and the small area data, it is determined what small areas the business parks are in and workplaces are randomly located in those small areas. Again using a combination of Google maps and the small area data, schools in the town are located and assigned to a random positions in the correct small area.

[1] Map data ©2016 Google

[2] `http://www.localbusinesspages.ie/`(Date Accessed 17/06/16)

3.2 Populating Agents

Agents are added to the town based on the 2011 census data. The number of agents in the simulation will be the number of people living in the town. The following steps are used in populating the town with agents and are performed for each individual small area:

- Each household is assigned a type *(single, couple, couple plus others, couple with children, couple with children plus others, single parent, single parent plus others or other)*.
- Adults are added into each household. One agent is added to households with types *single, single parent and other*. Two agents are added to the households with type *couple*.
- Adults in each household are assigned a sex and age based on a probability distribution determined from the CSO census age, sex tables for the relevant small area.
  - The age categories provided by the CSO are by year until 19 after which ages are reported in ranges of five years, for example ages 20-24 or ages 60-64, and then anyone over 85 is combined into one age bracket. For the purpose of the simulation the first year in the age range represents the entire range. For example, everyone in the 20-24 age range will be given the age of 20.
  - Couples are assigned opposite genders and ages within 10 years of each other [3].
- If a household type includes children a probability distribution determined from the relevant census data is used to determine if all children in the house are *under 15, over 15 or both over and under 15*.
- Data from the *family units with children by size and age of children* table is used to determine the probability that each household with type child has 1, 2, 3, 4 or 5 children in the household.
- Children are assigned a sex and age based on a probability distribution extracted from the relevant census data and the type of children the household is assigned *(under 15, over 15 or both under and over 15)*.
- If the total number of agents populating a small area is not equal to the total number of agents who should be in the area based on the CSO data, additional agents are added and randomly assigned to households of types *couple plus others, couple with children plus others, single parent plus others or other*.
- All agents are assigned an economic status based on CSO data.
  - Agents over 65 are assigned to *retired*.
  - Agents between the ages of 5 and 14 are assigned to *student*.
  - Agents between the ages of 15 and 18 are first assigned to *student*. If there are more agents aged 15 to 18 in the small area than the number of students in the same age categories then the agents are assigned to

[3] The 2011 Census does not record same-sex couples. This should, however, be expanded upon in future work

looking for first job. If there are still more agents aged 15 to 18 they are then assigned an economic status of *work, unemployed or sick/disabled* following the distribution for these categories for the relevant small area.
    - Adult agents under 65 are assigned to *work, looking for first job, unemployed, sick/disabled or stay at home* following the distribution for these categories for the relevant small area. Agents are only assigned to *stay at home* if they are part of a couple.
    - Agents under the age of 5 are assigned to *student* if they have no *stay at home* parent. If they have a *stay at home* parent then a probability determines if the agent will be assigned to *student*.
- Agents with an economic status of *work* are randomly assigned to one of the workplaces in town.
- Agents aged 13 and up with an economic status of *student* are assigned to a secondary school.
- Agents aged 4 to 12 with an economic status of *student* are assigned to a primary school.
- Agents aged 3 and below with an economic status of *student* are assigned to a preschool.

3.3 Transportation

In the model agents move between their home and destination in a straight line following the most direct route. Although this is a naive model, for small towns where distances travelled are short, such as those discussed in this paper, it is effective. In future work census data about modes of transport used could be used to drive more realistic transport model.

3.4 Schedule

Simulations within NetLogo use discrete time steps with agents' behaviours updated at each step. Each timestep in the model represents two hours in a day. Thus 12 timesteps represent one day, and 84 timesteps make up a week. Each time step travel is simulated in the model with agents moving between destinations. An agent does not move instantaneously to their new destination but moves in steps through the environment. This allows for contact with other agents to occur in passing and not just at the agent's start or end points. Agents' movements vary based on the economic status of the agent.

- Agents who are working leave their home on the fourth time step of the day which would be equivalent to between 8am and 9am, arrive at work over one time step, spend 4 time steps (8 hours) at work and then return home.
- Students also leave on the fourth time step but only spend 3 time steps (6 hours) at school.
- Stay at home agents who have children in primary school walk with their children to school on the 4th time step and then return home during the same time step. Between the fourth and seventh time step (when students

return home from school) stay at home agents move randomly throughout the town: at each step if an agent is at home they have a 50% chance of staying at home. If not at home an agent has a 50 % chance of picking a new destination in town and moving there. At the 7th time step of the day the stay at home agent will go to their child's school and then walk home.

- Stay at home agents who do not have a child in primary school move randomly throughout the town between the fourth time step and the seventh time step the same way stay at home parents move when they are not walking with their children to school or home.
- Agents younger than 4 who are not assigned to a preschool move with their stay at home parent throughout the day.
- Agents who are *unemployed, looking for their first job, retired or sick/disabled* move randomly throughout the town between the 4th and 10th time steps of the day.
- If an agent is infected with the disease simulated within a model then their behaviour is affected. Infected agents have a certain probability of staying home. If they are working the agents will stay home 30% of the time. Students will stay home 80% of the time. Unemployed agents will stay home 75% of the time, and stay at home agents with primary school children will stay at home 10% of the time when walking children to school and 50% of the time when moving around town. Stay at home agents with non-primary school children will stay at home 50% of the time.
- All agents will move randomly through the town on the weekends.

3.5 Disease Model

The disease model used in the simulation is based on an SIR model [21]. The SIR model categorizes individuals into susceptible, infected or recovered statuses and looks at movement of individuals between categories. Traditional SIR type models use ordinary differential equations to determine the rate of change of individuals in each category [13]. An agent-based model, however, uses a probability for between host transmission and within host progression. Between host transmission occurs when a susceptible agent comes into contact with an infected agent. At that time the susceptible agent will be infected or not based on a probability. Once an agent is infected with a virus the agent moves between infected and recovered based on parameters defined by the model [3]. In our model at each time step an agent will update a contact list of all agents that they have come into contact with during that time step. If a healthy agent comes into contact with an infected agent the healthy agent will have a certain probability of becoming infected. An infected agent has a probability at each time step that they will recover from the disease. The model is a general disease model which allows for adaption to multiple diseases.

4 Case Studies

We took two towns in Ireland to test the simulation, Rathnew and Bagenalstown. Both towns have roughly the same population but the area covered by the towns is different. Rathnew covers approximately 1 km^2 while Bagenalstown covers approximately 7 km^2.

4.1 Rathnew

Rathnew is a town in County Wicklow, Ireland with a population of approximately 2,964. It was chosen for modelling because it is comparable in size to the town modeled by Skvortsov et al. [21]. The town is located next to Wicklow town off the M11 motorway. There are three industrial areas in the town: Rathnew Business Park, Brommhall Business Park and the Village Mill Enterprise Park. There are approximately 126 businesses in the town including two hotels. There are no trains that stop in Rathnew but there is one local bus stop. Rathnew has three primary schools, one secondary school and three preschools.

Figure 2 shows a screenshot of the simulation of the town in NetLogo. Each agent is shown and the small area boundaries are highlighted. The inset shows a close up of one of the small areas to show the agents in the simulation.

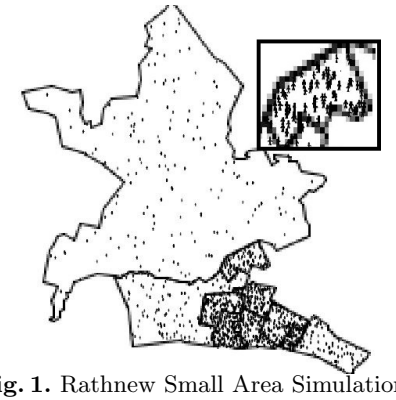

Fig. 1. Rathnew Small Area Simulation

The infection model used in the simulation is based on the infection model in Skvortosov et al. [21]. Initially five agents are infected. Every time a healthy agent comes into contact with an infected agent, the agent has a probability of 0.0071 of getting infected. Each time step every infected agent has a probability of 0.9959 of recovering. The infection and recovery probabilities are chosen to follow Skvortosov et al and model a flu-like disease [21].

The results from running the Rathnew model are shown in Figure 2. The graph shows how the percent of susceptible, infected and recovered individuals change over time. This is a standard infection curve and the way that epidemiological model results are reported. The curve matches the shape of the classic SIR infection curve [21]. This shows that the model is simulating an outbreak as expected.

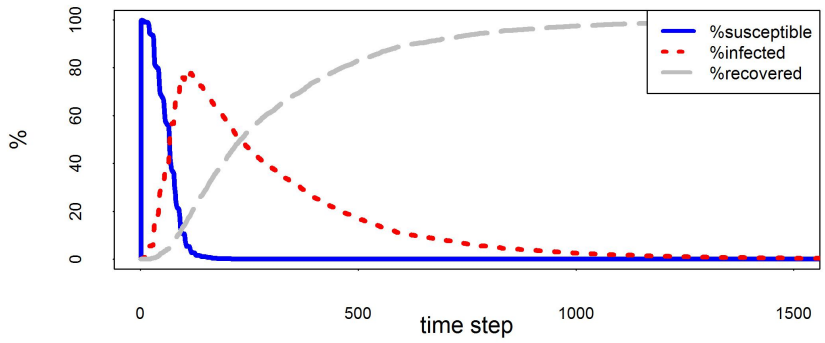

Fig. 2. The figure shows the infection curves from one run of a model of Rathnew with the percentage of agents susceptible, infected or recovered at each time step

4.2 Bagenalstown

Bagenalstown, also known as Muine Bheag and Muinebheag, is a town in County Carlow, Ireland with a population of approximately 2,950. It was chosen for modelling, as although it has a similar population to Rathnew, the population density is quite different. The town is located along the River Barrow at the junction of the R705 and R724 regional roads. There is one industrial area in the town. There is one train station in Bagenalstown. Bagenalstown has three primary schools, two secondary school and two preschools. Figure 3 shows a screenshot of the Bagenalstown simulation.

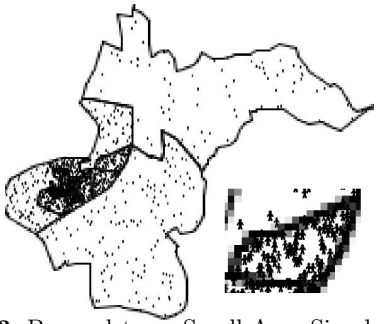

Fig. 3. Bagenalstown Small Area Simulation

The infection model used is the same as for Rathnew model. The results from running the Bagenalstown model are shown in Figure 4. Similar to the results from the Rathnew simulation, the curve matches the shape of the classic SIR infection curve [21].

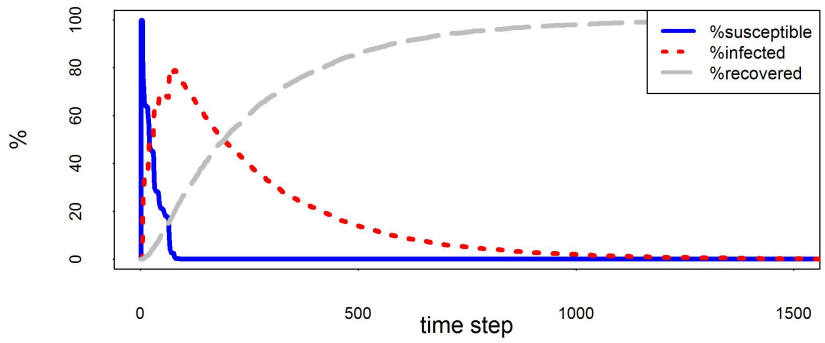

Fig. 4. The figure shows the infection curves from one run of a model of Bagenalstown with the percentage of agents susceptible, infected or recovered at each time step

5 Conclusion

The work described in this paper shows that it is possible to create a simulation model for a small town in Ireland that can be adapted to multiple towns using only publicly available data. The use of publicly available data is important in that it allows for easier access and reproducability. Only one epidemiological agent based model has been created for the population of Ireland and that model was not spatially explicity as ours is and it was was not designed to deal with a disease with airborne transmission as our model does [23]. The population in our model is an accurate representation of the age, sex, and family size breakdown of the actual towns being modelled. Being able to create an accurate simulation model for a town in Ireland allows for better results when adding the disease model into the simulation. Accurate smaller scale models can aid a community in deciding how best to deal with an outbreak, for example determining if and when to close schools. A smaller model would be less computationally intensive and thus take less time to run, allowing for faster results and results focused on a specific community instead of all of Ireland. However, the methods used in this simulation will be scaled up to model larger towns and cities, and will eventually be used to create a model of Ireland. Such a model could be used to help determine appropriate responses to future outbreaks including vaccine strategies, event cancellation or postponement policies on a national level.

Future work for the town model will include creating a more accurate transport model. In creating a better transport model the closed population concept will be evaluated and potentially altered. Although our model does not include commuting into and out of the town, these are potential sources of disease spread. Visitors could also be included in the model. Agents visiting the town can also

be a more realistic method of introducing a disease into the population this is how the 2016 measles outbreak in Ireland is believed to have started [17].

Better assignment to workplaces and schools could also be included. Instead of randomly choosing one of the age appropriate schools in town agents could be assigned to the school nearest to their home. Socio-economic status and education levels could also be used to group similar agents in workplaces. Currently in a model when agents are in the same place they are in contact with each other. For example, all students in a school are in contact for each time step they are there. This could be made more accurate by adjusting the model so that agents at the high density areas such as schools always come into contact with others in their social network but everyone else only at a certain probability.

A more sophisticated disease model could be created by including other factors such as vaccination status. Having a certain number of people in the town who are immune to the disease will change the infection dynamics and how it will spread. This will create a disease model that takes into account not only heterogeneous mixing but individual characteristics that could make someone more or less suceptible to a disease, allowing for more accurate results than a classic SIR model.

References

1. Data on individual schools. Department of Education of Education and Skills `http://www.education.ie/en/Publications/Statistics/Data-on-Individual-Schools/Data-on-Individual-Schools.html`
2. Ajelli, M., Goncalves, B., Balcan, D., Colizza, V., Hu, H., Ramasco, J.J., Merler, S., Vespignani, A.: Comparing large-scale computational approaches to epidemic modeling: Agent-based versus structured metapopulation models. BMC Infectious Diseases 10(190) (2010)
3. Barrett, C.L., Bisset, K.R., Eubank, S.G., Feng, X., Marathe, M.V.: Episimdemics: an efficient algorithm for simulating the spread of infectious disease over large realistic social networks. International Conference for High Performance Computing, Networking, Storage and Analysis (2008)
4. Bobashev, G.V., Goedecke, D.M., Yu, F., Epstein, J.M.: A hybrid epidemic model: Combining the advantages of agent-based and equation based-approaches. Proceedings of the 2007 Winter Simulation Conference pp. 1532–1537 (2007)
5. Crooks, A.T., Hailegiorgis, A.B.: An agent-based modeling approach applied to the spread of cholera. Environmental Modelling & Software 62, 164 – 177 (2014)
6. CSO: Census 2011 boundary files (2014), `http://www.cso.ie/en/census/census2011boundaryfiles/`, date accessed 26.05.2016
7. CSO: Census 2011 small area population statistics (saps) (2014), `http://www.cso.ie/en/census/census2011smallareapopulationstatisticssaps/`, date accessed 26.05.2016
8. Dibble, C., Feldman, P.G.: The geograph 3d computational laboratory: Network and terrain landscapes for repast. Journal of Artificial Societies and Social Simulation 7(1) (2004)
9. Dunham, J.B.: An agent-based spatially explicit epidemiological model in mason. Journal of Artificial Societies and Social Simulation 9(1), 3 (2005)

10. Friás-Martínez, E., Williamson, G., Friás-Martínez, V.: An agent-based model of epidemic spread using human mobility and social network information. IEEE Conference on Social Computing (2011)
11. Gilbert, N.: Agent-Based Models. 7-153, Sage Publications, Inc, London (2008)
12. Grimm, V., Berger, U., Bastiansen, F., Eliassen, S., Ginot, V., Giske, J., Goss-Custard, J., Grand, T., Heinz, S.K., Huse, G., Huth, A., Jepsen, J.U., Jørgensen, C., Mooij, W.M., Müller, B., Pe'er, G., Piou, C., Railsback, S.F., Robbins, A.M., Robbins, M.M., Rossmanith, E., Rüger, N., Strand, E., Souissi, S., Stillman, R.A., Vabø, R., Visser, U., DeAngelis, D.L.: A standard protocol for describing individual-based and agent-based models. Ecological Modelling 198(1–2), 115 – 126 (2006)
13. Keeling, M.J., Rohani, P.: Modeling Infectious Diseases in Humans and Animals. Princeton University Press, Princeton (2008)
14. Linard, C., Poncon, N., Fontenille, D., Lambin, E.F.: A multi-agent simulation to assess the risk of malaria re-emergence in southern france. Ecological Modelling 220(2), 160 – 174 (2009)
15. Mao, L.: Modeling triple-diffusions of infectious diseases, information, and preventive behaviors through a metropolitan social network—an agent-based simulation. Applied Geography 50, 31 – 39 (2014)
16. Olsen, J., Jepsen, M.R.: Human papillomavirus transmission and cost-effectiveness of introducing quadrivalent hpv vaccination in denmark. International Journal of Technology Assesment in Health Care 26(2), 183 – 191 (2010)
17. O'Regan, E., Larkin, L.: Measles passed on by patients in kerry outbreak (2016), http://www.independent.ie/irish-news/health/measles-passed-on-by-patients-in-kerry-outbreak-34826006.html, date accessed 08.07.2016
18. Perez, L., Dragicevic, S.: An agent-based approach for modeling dynamics of contagious disease spread. International Journal of Health Geographics 8(50), 1–17 (2009)
19. Rakowski, F., Gruziel, M., Bieniasz−Krzywiec, L., Radomski, J.P.: Influenza epidemic spread simulation for poland — a large scale, individual based model study. Physica A: Statistical Mechanics and its Applications 389(16), 3149 – 3165 (2010a)
20. Rakowski, F., Gruziel, M., Krych, M., Radomski, J.P.: Large scale daily contacts and mobility model - an individual-based countrywide simulation study for poland. Journal of Artificial Societies and Social Simulation 13(1) (2010b)
21. Skvortsov, A.T., Connell, R.B., Dawson, P.D., Gailis, R.M.: Epidemic modelling: Validation of agent-based simulation by using simple mathematical models. International Congress on Simulation and Modelling pp. 657–662 (2007)
22. Team, Q.D.: Qgis geographic information system 2.8. Open Source Geospatial Foundation (2009), http://www.qgis.org/en/site/index.html
23. Usher, C., Tilson, L., Olsen, J., Jepsen, M., Walsh, C., Barry, M.: Cost-effectiveness of human papillomavirus vaccine in reducing the risk of cervical cancer in ireland due to {HPV} types 16 and 18 using a transmission dynamic model. Vaccine 26(44), 5654 – 5661 (2008), http://www.sciencedirect.com/science/article/pii/S0264410X08010128
24. Wilensky, U.: Netlogo. Center for Connected Learning and Computer-Based Modeling, Northwestern University. Evanston, IL. (1999), https://ccl.northwestern.edu/netlogo/