AN EPIDEMIOLOGIC MODELING AND DATA INTEGRATION FRAMEWORK

Pfeifer B[1], Seger M[1], Netzer M[1], Osl M[1], Modre-Osprian R[2], Schreier G[2], Hanser F[3], Baumgartner C[1]

Abstract

In this work a cellular automaton software package for simulating different infectious diseases, storing the simulation results in a data warehouse system, and analyzing the obtained results in order to generate predictions models or contingency plans is proposed. The Brisbane H3N2 flu virus, which has been spreading this winter season, was used for spreading simulation in federal state of Tyrol. The simulation-modeling framework consists of an underlying cellular automaton. The generated data are stored in the back room of the data warehouse using the Talend Open Studio software package, and subsequent statistical and data mining tasks are performed using the Knowledge Discovery in Database Designer (KD^3). The obtained simulation results were used for generating prediction models for all nine federal States of Austria, and were compared to flu outbreaks in the past.

1. Introduction

In the field of model building and simulation in medical informatics, methods and tools are developed for acquiring data, which are then processed and analyzed using bio statistical and data mining methods. By applying these methods new knowledge is generated that helps contributes towards a better understanding of the underlying processes, which enables to generate and improve existing strategies, treatments, and workflows.

Communicable or infectious diseases kill more people worldwide than any other single cause. Pathogenic agents such as bacteria, viruses, parasites or fungi are responsible for those diseases. The disease transmission from individual to individual occurs by physical contact with an infected individual, usage and handling of contaminated substances like food and liquids, airborne inhalation, body fluids, vectored infection to name just a few.

When studying historical documents some endemic and/or pandemic outbreaks are described. In the years from 1347 to 1352 over 25 million Europeans (about one third of Europe's population at this time) were killed by the plague, which is better known as "black death". In 1896 the next outbreak of the plague occurred, and spread to nearly every part of the globe, and killed about 12 mil-

[1] UMIT, Institute of Biomedical Engineering, Eduard Wallnöfer Centre 1, 6060 Hall/Tirol
[2] ARC Seibersdorf research GmbH, Innsbruck, Austria
[3] UMIT, Research Division for Pervasive Health, Eduard Wallnöfer Centre 1, 6060 Hall/Tirol

lion individuals up to 1945. The Spanish Flu outbreak occurred in the years between 1918 and 1920 and was able to kill between 20 and 50 million people, especially young adults and teens with well working immune systems. In 1957 about one million individuals fell prey to the Asian Flu, and the Hong Kong Flu killed about 700.000 individuals. The Acquired Immune Deficiency Syndrome (AIDS), which is caused by the immunodeficiency virus (HIV), was first recognized in the 1980s. Its death toll lies by about 20 million individuals. The HI-Virus infects about 50 million individuals all over the world. More historical data can be obtained at [1].

This shows that infectious diseases and the generation of contingency plans and strategies are of importance to check an outbreak, which means to save lives and to minimize economical impact.

As every year the flu (especially during winter season) afflicts Austria, as well as other countries. During the Christmas time the Influenza type A subtype H3N2 virus, named "Brisbane", has exponentially been spreading [2]. Typically, about 10 percent of adults and 15 percent of children fall ill. The symptoms are: sudden high fever, shooting pains and arthralgia.

For modeling different communicable diseases several approaches were presented in the past [3-5]. Our proposed framework enables to simulate the spatio-temporal individual based behavior of different communicable diseases and to persistently store the simulation results for further analysis using data mining techniques.

2. Methods

2. 1. Infectious Disease Simulation

The classic epidemic model, which is known as S-I-R model, is widely used for simulating the behavior of communicable diseases [3-5]. In the S-I-R model the class S denoted the number of susceptibles, class I denotes the number of infectives, and class R refers to the number of recovered people in the model. The model is given by the initial value problem:

$$\frac{dS}{dt} = -aIS, \quad S_0 \geq 0$$
$$\frac{dI}{dt} = aIS - bI, \quad I_0 \geq 0$$
$$\frac{dR}{dt} = bI, \quad R_0 \geq 0$$

where the sum of the model classes S(t)+I(t)+R(t)=N, with N being the number of all individuals.

However, the classic SIR model is unable to take account of natural birth and death, immigration and emigration, passive immunity and spatial arrangement. Therefore, partial differential equations (PDE) can be used. Following PDE can be used for modeling infection diffusion through space:

$$\frac{dS}{dt} = -(aI + \beta(\frac{\partial^2 I}{\partial x^2} + \frac{\partial^2 I}{\partial y^2}))S, \quad S_0 \geq 0$$

$$\frac{dI}{dt} = (aI + \beta(\frac{\partial^2 I}{\partial x^2} + \frac{\partial^2 I}{\partial y^2})) - bI, \quad I_0 \geq 0$$

$$\frac{dR}{dt} = bI, \quad R_0 \geq 0$$

With these models it is possible to simulate the spreading of a disease over a population in space and time, but if tracking of individuals or geographical conditions and demographic realities are needed these models are also inadequate, and therefore, cellular automata (CA) modeling approach can be used.

For this purpose cellular automata (CA) can be used. A CA is a dynamical system in which time and space is discrete; it is specified by a regular discrete lattice of cells and boundary conditions, a finite set of cells and states, a defined neighborhood, and a state transition function for cell evolution over time. A CA can be formal described as a 9-tuple

$$CA = (C, N, \sum, \Psi, Q, \delta, \sigma, q_0, F),$$

where C is the cell adjustment, N is the neighborhood definition, $\sum$ denotes the input alphabet, Ψ the output alphabet, and Q describes the set of internal cell states. δ denotes the state transition function with $\delta: Q \times Q^{n_i} \rightarrow Q$, and σ is the output function with $\sigma: Q \rightarrow \Psi$. q_0 denotes the initial state with $q_0 \in Q$, and F denotes the finites states with $F \subseteq Q$, respectively.

The disease modeling and simulation framework was implemented using Java SE 1.6. As object oriented paradigms and software patterns were used, the software package is highly generic, which means, that approaches based on cellular automata can be easily integrated and implemented using this framework. Several classes exist in the kernel of the simulation and modeling package. Only the two generic CA package classes are described here briefly. The class `CellularAutomaton` defines a generic cellular automaton and provides the abstract method `compute()`, which is responsible for performing one simulation step over the cells. The class `Cell` represents one generic cell with the associated properties and neighborhood relation. The abstract method `performCellAction()` has to be overridden for performing the cell state transition function. Furthermore, the resulting states and attributes are persistently stored for each cell and each individual for further analysis; for each time step also a snapshot of the geographical map is generated and stored as portable network graphics (PNG) file.

2. 2. Data Integration

A data warehouse is defined as subject oriented, integrated, non-volatile collection of integrated data [6]. The generated life sciences data need to be stored in a data warehouse in order to perform further analysis tasks. The underlying schema used in the backroom component is the so-called star schema. This schema is characterized by one or more large fact tables that contain a number of the primary information in the data warehouse, and a number of much smaller dimension tables, each of which contains information about the entries for a particular attribute in the fact table. A query on such a schema is characterized as a join between fact tables and several dimensional tables.

For the data integration tasks the freely available open source software package provided by Talend Open Data Solution is used [7]. The Talend Open Studio enables the backroom specialist to develop data extraction, transformation (cleansing) and loading (ETL) components. The software then generates Java code that can be used in own software solutions. Furthermore, it is possible to implement own components, if the offered components do not fulfill the requirements. *Figure 1* shows the graphical user interface of the Talend Open Studio design tool.

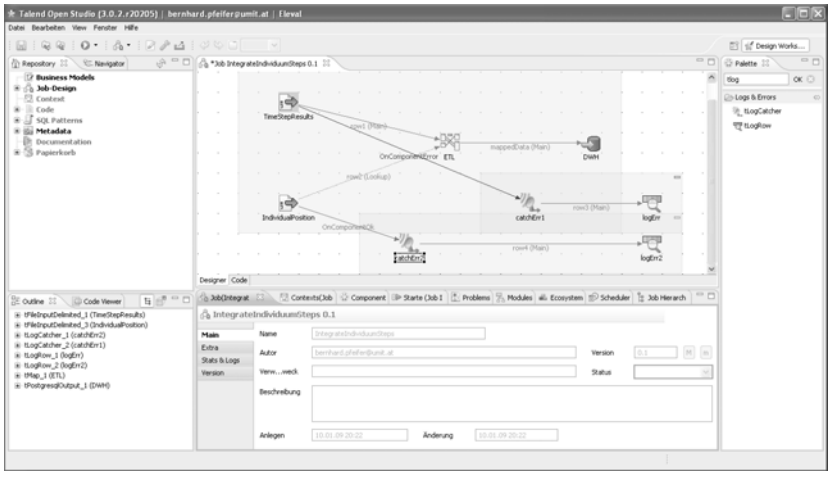

Figure 1: Talend Open Studio Designer, which was used for integrating the generated simulation data. For more information see text.

2. 3. Knowledge Discovery in Simulation Data

When performing knowledge discovery several steps are usually involved. These steps incorporate focusing and describing the problem, preprocessing and transformation of the data, performing statistical analyses and data mining algorithms, and a concluding evaluation [8]. As all the simulation result data is stored in the data warehouse system the preprocessing and transformation tasks can be reduced. So, one can focus on building analyzing workflows in order to generate new knowledge. Therefore, using the KD^3 (Knowledge Discovery in Database Designer) [9] software package the user can focus on building analyzing workflows to interpret the simulation results. Such a workflow is characterized by a repeatable pattern activity, which is used for solving defined processes using different parameters and data sets in simulation scenarios. So-called functional objects (FOs) are the key elements of the application. A FO consists of in- and out ports, where the data set can be transported from one component to another. By implementing new FOs the KD^3 application can be easily expanded by new functionality. The new objects are then loaded into the KD^3 workspace by using the Java Reflection API. *Figure 2* depicts the KD3 designer application that is implemented for data mining and analyzing the Brisbane flu simulation.

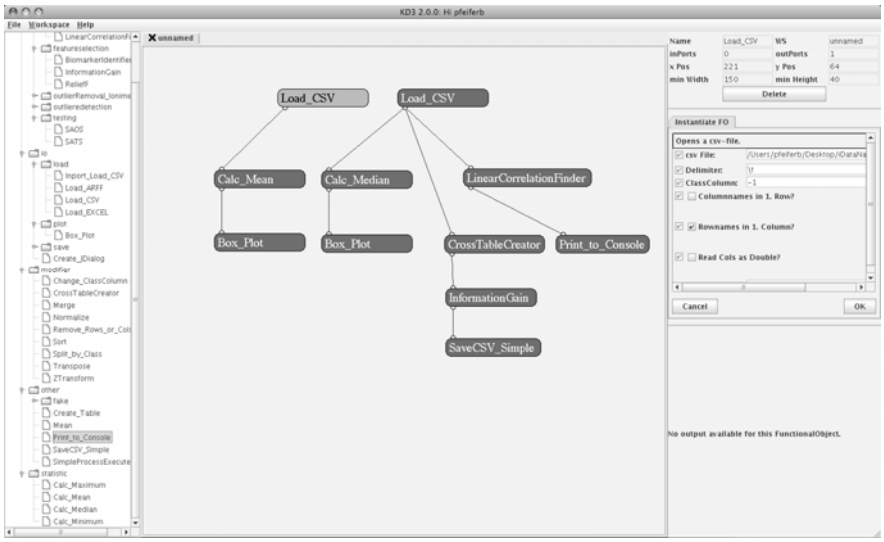

Figure 2: Screenshot of the knowledge-mining task of the Brisbane flu simulation results.

3. Results

In this study the Brisbane flu was simulated for the federal state of Tyrol. The seed point of the infection was set to the capital Innsbruck. Furthermore, using two different drugs medical treatment of the infected individuals was performed. Medication 1one is able to reduce lethality and severe symptoms by 55 percent, and medication two by 45 percent. Furthermore, the social behavior of the individuals changes during the disease spread, which means that the individuals try to avoid unnecessary social contact. This would also occur in real situations.

The run time of the simulation was 38 minutes for 365 simulated days. For running the simulation an Apple X-Serve 1.1 OS X Server Version 10.4.10 with 2x2GHz Dual Core Intel Xeon processor and 2GB of RAM was used.

The simulation data was integrated using the Talend Software and analyzed with the KD^3 workflow mining tool. Although, only the federal state of Tyrol was simulated it is possible to give a prediction for the other nine federal states of Austria. The simulation results show that approximately 240 fatal outcomes could occur in Tyrol, which is 8% of the total fatal outcome that could occur in Austria. *Figure 3a* depicts the percentages of the fatal outcomes in the nine Austrian states. Furthermore, the number of possible infections for the federal state of Tyrol was estimated to be 44.787 individuals. The infection prediction for the other states is depicted in *Figure 3b*.

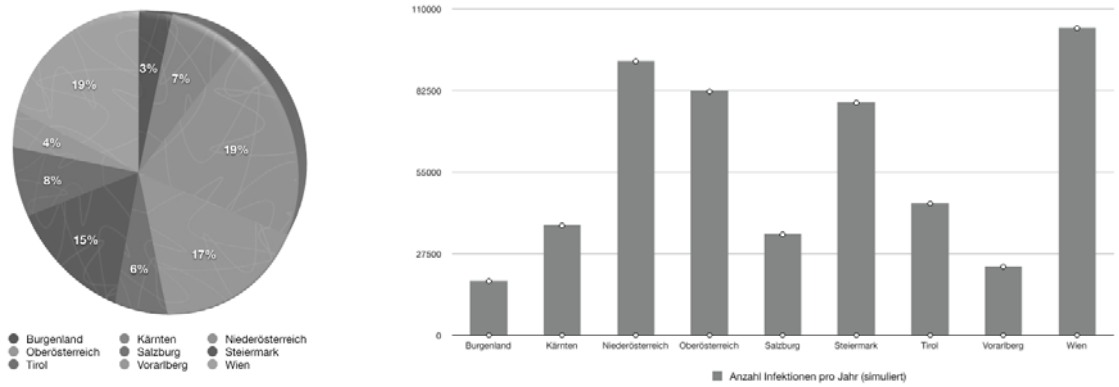

Figure 3: (a) Percentage of the estimated outcomes in all nine federal states of Austria; (b) Estimated infections for the nine states of Austria; for more information see text.

Figure 4 depicts the different states of the possible individual classes. During the first 140 days most individuals get infected, after that initial infection period individuals become susceptible again, which means that the disease might infect them again. Some of the individuals get permanently immune against the disease.

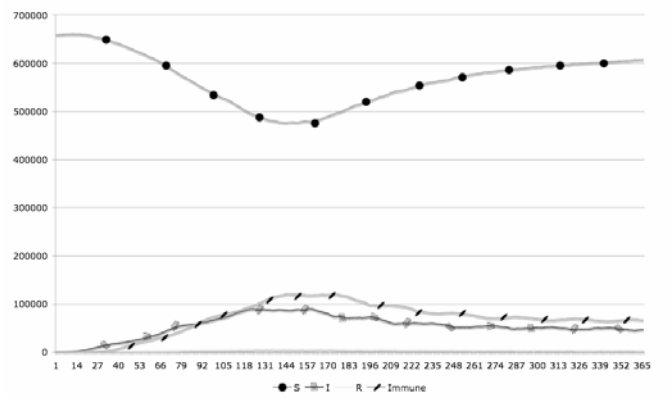

Figure 4: The sphered line (S) is used for the number of susceptibles in the population, (I) shows infected individuals, (Immune) shows immune individuals and the (R) line is used for depicting "removed" individuals.

Figure 5 depicts the spatial arrangement of the disease spreading. The black markers used in *Figure 5* demonstrate, where infected individuals are located and how the infection spreads over the population in space and two distinct time instances.

Figure 5: Population density map of the federal state of Tyrol with superimposed infection map generated by the simulation run. The left figure shows the situation 20 days after the infection has started. The right figure indicates the spreading situation after 120 days. It can be can seen that after 120 days nearly every part of federal state Tyrol is involved, but the infection density is much lower compared to the beginning of the infection.

4. Discussion and Outlook

The framework enables the simulation of different infectious diseases. All the generated data is subsequent stored in a data warehouse system, which enables to perform analytical tasks for generating prediction models or improvement of contingency plans. Furthermore, our simulations are in close agreement to past observations and records [10,11].
Public health offices could use the tool for getting deeper understanding in the spreading mechanisms and for preventing serious long-term economic repercussions.

5. Acknowledgement

We thank the Federal Ministry of Economics and Labour of the Republic of Austria, the Tyrolean Future Foundation and the Competence centre HITT-health Information Technologies Tyrol for funding this work.

6. References

[1] WORLD HEALTH ORGANIZATION (WHO). Available at: http://www.who.int

[2] Austria to face flu epidemic in new year (2009);
http://www.wienerzeitung.at/DesktopDefault.aspx?TabID=4975&Alias=wzo&cob=388656

[3] K. LICHTENBERGER. Stochastic cellular automaton models in disease spreading and ecology. Oct 2005. Doctoral Thesis, TU-Graz

[4] HW. HETHECOTE. The mathematics of infectious diseases. Society for Industrie and Applied Mathematics. Vol 42, pp.599-653, Oct 2000.

[5] S. EUBANK, H. GUCLU, V. A. KUMAR, M. MARATHE et al., “Modelling disease outbreaks in realistic urban social networks,” Nature , Jan 2004.

[6] R. KIMBALL AND J. CASERTA, The Data Warehouse ETL Toolkit, Wiley Publishing Inc., 2004.

[7] TALEND OPEN STUDIO. www.talend.com, 2008.

[8] U. M. FAYYAD, G. PIATETSKY-SHAPIRO, AND P. SMYTH, “From data mining to knowledge discovery: An overview,” in Advances in Knowledge Discovery and Data Mining, pp. 1–34, 1996.

[9] PFEIFER B, TEJADA MM, KUGLER K, OSL M, NETZER M, SEGER M, MODRE-OSPRIAN R, SCHREIER G, TILG B. A Biomedical Knowledge Discovery in Databases Design Tool - Turning Data into Information. eHealth 2008, Wien, 29.-30. Mai 2008.

[10] DIAGNOSTISCHES INFLUENZA NETZWERK ÖSTERREICH: http://www.influenza.at/influenza-main.htm

[11] KÄRNTNER GEBIETSKRANKENKASSE. Influenza eine nicht ungefährliche Krankheit: http://www.kgkk.at/portal/index.html;jsessionid=C7A0928475E0E293FF9CC58B5BF544C7?ctrl:cmd=render&ctrl:window=kgkkportal.channel_content.cmsWindow&p_menuid=7602&p_tabid=2&p_pubid=11773