**Noname manuscript No.**
(will be inserted by the editor)

# Improving process algebra model structure and parameters in infectious disease epidemiology through data mining

**Dalila Hamami, Baghdad Atmani, Ross Cameron, Kevin G Pollock, Carron Shankland**

Received: date / Accepted: date

**Abstract** Computational models are increasingly used to assist decision-making in public health epidemiology, but achieving the best model is a complex task due to the interaction of many components and variability of parameter values causing radically different dynamics. The modelling process can be enhanced through the use of data mining techniques. Here, we demonstrate this by applying association rules and clustering techniques to two stages of modelling: identifying pertinent structures in the initial model creation stage, and choosing optimal parameters to match that model to observed data. This is illustrated through application to the study of the circulating mumps virus in Scotland, 2004-2015.

**Keywords** epidemiological modeling · mumps infection · process algebras · Bio-PEPA formalism · data mining · association rules · clustering · time series.

Dalila Hamami
University of Ahmed Benbella Oran 1, Algeria,
E-mail: dhamami8@gmail.com

Baghdad Atmani
University of Ahmed Benbella Oran 1, Algeria,
E-mail: atmani.baghdad@gmail.com

Ross Cameron
Health Protection Scotland, UK.
E-mail: ross.cameron@nhs.net

Kevin G Pollock
Health Protection Scotland, UK.
E-mail: kevin.pollock@nhs.net

Carron Shankland
University of Stirling, UK.
E-mail: ces@cs.stir.ac.uk

Accepted for publication in *Journal of Intelligent Information Systems*. The final publication is available at Springer via https://doi.org/10.1007/s10844-017-0476-1

## 1 Introduction

Epidemiological systems are defined by behaviour which may depend on a multitude of events occurring in space and over time. From the initial population models of Malthus [1] and the classic epidemiological models of Anderson and May [2] much effort has been expended in developing more accurate mathematical, and latterly computational, models of epidemics. These mechanistic models have been used successfully in a variety of ways to explore disease spread and control and thus to inform decision-making, see for example [2–6], but their effective use depends critically on getting the right model in the first place and correctly parameterising that model. For instance, a small change to either the structure of the model, or to a parameter value, can make the difference between the model predicting a large infectious disease outbreak or disease elimination [7]. Although analysis of models is supported by tools, until recently creation of models has been done entirely by hand, relying on the ingenuity of the modellers and good communication with epidemiologists to understand system behaviours. In contrast, techniques from data mining, machine learning, and evolutionary algorithms [8–10] can be used to process observed data and produce predictive models. Such models are implicit, 'black box' models: while they can be used to predict future behaviour, they have no power to explain the underlying processes. We propose harnessing machine learning and data mining techniques to assist the craft of mechanistic model construction.

In this paper, we explore two particular data mining techniques: association rules and clustering, and how these can be used both to identify pertinent structure in the development of the model, and to resolve the problem of parameter value identification. Both techniques are designed to draw out relationships, either between attributes of the data set (association rules and clustering), or between instances of data (clustering). This information can be used by an expert modeller to help them manually formulate a suitable model. We illustrate the use of data mining for modelling by developing an epidemiological model for the mumps virus using the process algebra formalism Bio-PEPA (Bio-Performance Evaluation Process Algebra) [11]. The primary advantage of Bio-PEPA is that an explicit model is obtained, which has a clear modular specification style and which gives access to different analyses for a whole population: stochastic simulation, model checking and ODE-based analyses [11–13]. Bio-PEPA simulations can easily produce time series data; however, constructing the model, choosing the right components, and identifying the optimal range of parameters to match observed data can require considerable expertise and time. In contrast, data mining provides individual-level information through analysis of observed data to extract key model features, and analysis of complex and varied time series outputs for a range of parameter settings. The combination of process algebra and data mining is only beginning to be explored [14,15] and plays to the complementary strengths of each.

The rest of the paper is organized as follows: in the next section, we provide a brief overview of related work in the area of model inference for process algebra and parameter estimation using data mining for infectious disease epidemiology. Section 3 describes our novel combination of clustering techniques and association rules with Bio-PEPA. The approach is illustrated through an application to mumps, assisting in both model structure and parameter identification, showing how a model which fits well to observed data is obtained. Mumps was chosen because current data shows interesting dynamics and data mining can help to explain those dynamics. Finally, we conclude by summarizing the achieved goals, discussing the outcomes and providing some directions for future work.

## 2 Related work and Background

Application of machine learning and data mining techniques to biological systems has grown considerably in recent years [16,17] but the launch editorial of BioData Mining [18] highlights the need to apply these techniques specifically in the area of epidemiology to create better predictive models. We consider works in building model structure in process algebra, and in parameter analysis using data mining, including clustering for time series analysis.

Existing work in model structure inference for process algebra lies mainly in the genetic programming area: process algebra models have been reconstructed given a partial initial model and target data by Ross and Imada [19] (structure only) and Oaken et al [20] (both structure and parameters). Neither method has so far been tested with a complex system and observed data. The approach of Bartocci et al [21] deals with noisy observed data in moderately complex systems, deriving high-level temporal logic specifications of systems (structure). Hamami and Atmani [15] combined association rules with process algebra to extract new pertinent rules with which to manually refine the tuberculosis model structure. Here, we augment this process with clustering, which is a novel combination. There are no works specifically relating clustering to Bio-PEPA or process algebra.

Much more work has been done in the area of parameter identification. Sumner [22] argues that model parameters are often estimated in a large range of values or associated with a high level of uncertainty. Consequently, this leads to a low confidence in simulation results of models and potential bias in parameter choice [23,24]. Several works have pinpointed sensitivity analysis as a solution to this problem [25,26]. However, sensitivity analysis quantifies the response of model output variables to parameter variation within a selected parameter space. It does not generate parameter values and can only be used to confirm optimal parameters.

Georgoulas et al [27] apply quantitative generalisation of Constraint Markov Chains to determine parameters primarily for Markov Jump processes (a close semantic relative of process algebra). Their focus is on obtaining statistical distributions concerning optimal parameters for the model when matching

to observed data. The Bayesian approach to parameter optimisation is also adopted by Bortolussi et al [28] and Bartocci et al [21,29]. In these works the model specification is given by temporal logic. This is not sufficiently explanatory or mechanistic for our purposes. The Oaken study [20] mentioned above generates optimal parameter sets and statistical distributions using genetic algorithms. This can be done simultaneously with structure inference. An advantage of all of these approaches is that statistical distributions of parameter values are generated.

Clustering is also used here on time series data output from the model. This is a fairly well-understood application area: the survey of Liao [30] presents a variety of approaches, some of which can be applied to group time series together in this way, and some of which are designed to identify sub-sequences within a time series. Our approach identifies model structure and optimal parameter settings for a process algebra model, based on a novel combination of clustering with association rules applied observed data, and of clustering applied to a large sample of model outputs.

### 2.1 Bio-PEPA formalism

Bio-PEPA (Bio-Performance Evaluation Process Algebra) is a formal language developed by Ciocchetta and Hillston [11] to describe biological systems. Bio-PEPA is a formalism allowing the description of a compartmental model of a system composed of interacting individuals. Models are based on a set of species (entities) and their dynamics, described by a set of actions and kinetic rules. This is similar to the style of compartmental models in Ordinary Differential Equations (ODE), e.g. Anderson and May [2] for epidemiology. The advantage of Bio-PEPA over ODE is the ability to analyse the model in different ways (deterministic or stochastic simulation, model checking, simulation traces), implemented through the Bio-PEPA plugin [11]. Models in Bio-PEPA are described by the following formal syntax, taken from [31]:

$$
\begin{aligned}
S &::= (\alpha, \kappa)\ op\ S \mid S + S \mid C \\
&\qquad \text{where } op\ = << \mid >> \mid (+) \mid (-) \mid (.) \\
P &::= P \bowtie P \mid S(x)
\end{aligned}
$$

where $S$: species corresponding to the main agents of the model. $\alpha$: actions, as chosen by the modeller, $\kappa$: stoichiometric coefficients of those actions, and $op$: operations describing the dynamics of $S$ (change up, change down, or influence on other actions). In the Species definition, $+$ allows $S$ to undertake a choice of different behaviours driven by the competing rates of the actions available. The operator $<<$ (resp. $>>$) indicates that the level of species increases by their stoichiometry coefficient (resp. decreases). The operators $(+)$, $(-)$ and $(.)$ indicate an activator, an inhibitor and generic modifier respectively. The latter is neither an activator nor an inhibitor and indicates the species is involved in the reaction where its level remains unchanged. In the Model definition ($P$),

models can be combined in parallel, with or without communication, where $x$ defines the initial population size.

Bio-PEPA has been used widely for epidemiological modelling [31–35]. Ciocchetta and Hillston [31] developed the first epidemiological Bio-PEPA model that was applied to avian influenza virus. The model was able to deal with population-level dynamics of a well-mixed group of individuals with specified attributes, stochasticity and spatial structure. Benkirane et al [32] developed a measles virus model, in which seasonality and immigration affected spread of disease. Bio-PEPA was shown to provide easy-to-construct, simple models by Hamami and Atmani [33,34] for existing models of both herpes zoster [36] and tuberculosis [37]. Another advantage of Bio-PEPA over some other modelling techniques and computer programming is its compact formal syntax: this has been shown useful when combining process algebra models with other techniques to investigate parameter values and model structure. For example, Ramanathan et al [35] used metamorphic testing and visualization with Bio-PEPA to study how the main parameters of epidemics (transmission, infection, demographics) affect model dynamics. Evolutionary techniques combined with Bio-PEPA were used by Oaken [20] to propose refined epidemiological models.

## 2.2 Data mining

It is beyond the scope of this work to present a complete review of data mining: see, for example Pardalos et al [16] and Sullivan [38] for reviews of data mining specifically applied to biomedical and life sciences. Instead, we summarise the pertinent features of the two chosen techniques: association rule learning and clustering.

Witten [39] defines association rule learning as the process which provides a set of rules able to express the relationship between a group of attributes appearing frequently together in large datasets. Formally, the rule is defined by a combination of a set of items in the form: $X \Rightarrow Y$, where $X$ is the antecedant and $Y$ the consequent. To evaluate the quality of mined rules, two main metrics are used: support and confidence. Assume a dataset consisting of transactions (rows). Support is the frequency of items appearing in the dataset. Support of $X \Rightarrow Y : P(X, Y)$ = (set of transactions containing both $X$ and $Y$) / (total set of transactions). Confidence evaluates reliability: it is the number of times that the rule has been found true. Confidence of $X \Rightarrow Y : P(Y|X)$ = (set of transactions containing both $X$ and $Y$) / (set of transactions containing $X$).

For example, outbreaks of childhood diseases are often correlated with school terms. Deriving association rules for such a dataset should produce rules which link week of the year to occurrence of the disease, with high confidence and support.

Association rules algorithms have been widely used in the healthcare field, as highlighted in the survey by Tomar and Agarwal [40]. The most commonly

**Table 1** Association rules and Clustering algorithms (taken from [39])

Association rules

| Algorithm | Function |
|---|---|
| Apriori | Derives association rules identifying general trends in the data. |
| Predictive Apriori | An extension of Apriori which balances support with confidence. |
| Tertius | Derives association rules by confirmation of first-order logical clauses. |

Clustering

| Algorithm | Function |
|---|---|
| Cobweb | Implements the Cobweb and Classit clustering algorithms |
| DBScan | Nearest-neighbor-based clustering that automatically determines the number of clusters |
| EM | Cluster using expectation maximization |
| FarthestFirst | Cluster using the farthest first traversal algorithm |
| HierarchicalClusterer | Agglomerative hierarchical clustering |
| sIB Cluster | Cluster using the sequential information bottleneck algorithm |
| SimpleKMeans | Cluster using the k-means method |
| XMeans | Extension of k-means |

reported association rules algorithms in the literature [39, 40] are: Apriori, Predictive Apriori and Tertius.

Clustering is a widely-used data mining technique in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics. Witten [39] describes clustering as the task of partitioning a set of objects so that objects in the same group (or cluster) are more similar to each other by some chosen measure than they are to those in other groups. The results can be used to generate hypotheses about important data features, to aid in visualization, or to reduce the data to a few representative points. Witten classifies these unsupervised learning techniques into four groups:

- Exclusive: each instance belongs to one and only one cluster.
- Overlapped: an instance could belong to several clusters.
- Probabilistic: an instance belongs to each cluster with a specific probability.
- Hierarchical: an instance is assigned to a cluster according to a hierarchical structure.

Clustering can be used on its own, as above, or as a prior step to association rule mining. In the latter case, use of clustering enhances association rule performance evaluation such as support and confidence.

## 3 Approach and Data Mining Applied to Mumps

The aim of this study is to improve computational modelling through combination with data mining techniques. The particular goal is to assist modellers in creating a suitable computational model with optimal parameter settings

to match observed data. This will allow improved outbreak prediction. There are many formalisms for computational modelling and several different approaches to data mining. Here, we pair Bio-PEPA with association rules and clustering, for reasons outlined above. We apply the approach to mumps data from Scotland.

### 3.1 Mumps prevalence in Scotland

Mumps is caused by a virus in the Paramyxoviridae family. Mumps, which affects only humans, is often a self-limited infection but may in some cases cause complications requiring hospitalization or even leave long-term side-effects. Mumps is prevalent throughout the world, despite widespread measles-mumps-rubella (MMR) vaccination programmes. The first dose of the trivalent MMR was introduced in Scotland in 1988 [41,42]. Scotland experienced a large national outbreak in 2015 with over 800 laboratory-confirmed cases [43] despite high MMR vaccine uptake of 95%. Health Protection Scotland (HPS) have provided us with data from 2004-2015 indicating that mumps occurs each year with a mix of major and minor epidemic waves (see Fig. 1). It is therefore not clear how the disease will spread in future. Are the large outbreaks outliers in a steady decline of mumps or do they signal a resurgence of mumps? Why was there only temporary elimination of the disease following widespread vaccination? What is causing mumps to persist? Producing a well-parameterised model which fits this data will inform healthcare strategies by helping epidemiologists to understand the underlying dynamics of the disease, predict the likely pattern of future outbreaks, and propose additional control measures such as supplementary vaccine efforts for student populations.

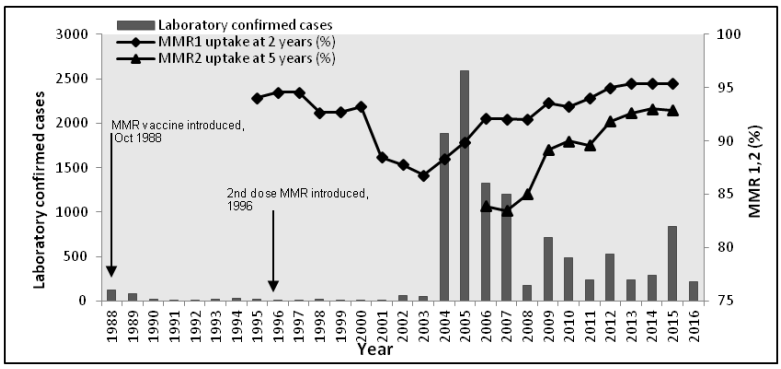


**Fig. 1** Mumps confirmed cases, Scotland 1988-2016 and MMR vaccine uptake

### 3.2 Observed data

Daily-reported data concerning mumps cases in Scotland from 2004 to 2015 is defined by the attributes Age, Sex, NHSBoard, Year, Week, Report Date, Disease. Data is pre-processed to replace unknown values with the median value of the range, and discretised to group age and weeks. The attribute Disease is removed as all cases have mumps and this adds no information. Aggregate vaccine status from an outbreak in a particular health board region in one year revealed that vaccination status might be an important feature in outbreaks. This is counter-intuitive as, for example, vaccination of measles confers life-long immunity. While we were unable to obtain observed data at this level of detail, we are able to construct a suitable simulated data set (OneYearOneBoardVaccStatus) manually combining observed cases with projected vaccination status, based on that outbreak, where 50.4% were fully vaccinated, 18.5% partially vaccinated, 12.6% were unvaccinated and 18.5% with unknown vaccine status [42]. Thus, the constructed data set (OneYearOneBoardVaccStatus) describes daily-reported individuals infected by mumps where each line (instance) corresponds to a reported individual and columns define the attributes describing the individual by Age, Sex, NHSBoard, Year, Week, Report Date and vaccine status.

A number of additional data mining experiments were carried out, combining additional data sources with this data (weather, transport, population density, immigration and emigration); however, none of these was found to provide significant correlation with epidemic outbreaks.

### 3.3 Outline Method

Clustering is used firstly as a prior step to association rules processing to maximize the identified features and then enhance the model structure. Secondly, it is used here as an analysis tool to identify optimal parameters for that model, combining clustering with time series. Our general approach is structured as follows:

1. Cluster on reported cases (observed data).
2. Extract pertinent features from clustered data using association rules algorithms.
3. Construct an enhanced Bio-PEPA model (manually) using the pertinent features.
4. Cluster on simulated data (model outputs) to select the best fit simulation to observed data.
5. Tune model parameters according to best fit simulation.

Thus, a range of association rule learners and clustering algorithms are used. These are summarised in Table 1 and are provided by the WEKA (Waikato Environment for Knowledge Analysis) [44]. It is not our goal to select one of these as better than the others for the task but to illustrate the use of readily

available algorithms in our setting and to analyse the outputs for additional insight to our epidemic system. In our experiments the default settings of WEKA are used for each algorithm.

In the following, the steps above are described in more detail, followed by application to the mumps example. Section 3.4 describes steps 1 and 2. Section 3.5 describes step 3. Section 3.6 covers steps 4 and 5.

## 3.4 Data Mining for model structure

Modelling is done by hand, based on the SEIR model template. For mumps the SEIR compartmental model is a commonly used starting point [2], where S: an individual who is susceptible to acquiring the virus, E: an exposed individual who has acquired the virus but is not yet infectious, I: an infectious individual and R: a recovered individual who has acquired immunity for life. Normally, adjustment of this basic template would be based on interaction between the domain expert and the modelling expert to reach a shared understanding of the system. To shortcut the modelling process and assist the modeller, we propose the use of data mining. Data mining might suggest additional pertinent features which will be used to enhance the model by restructuring with new compartments, adding new functional rates and recalculating parameters. Of course, this must still be validated through simulation and consultation with the domain expert, but the aim is to make this process more efficient. Clustering is used to group observed data related to infected cases, drawing out similarities between individuals, providing a way of identifying important system features. To understand and identify the common features of those clustered individuals, association rules are applied.

The rationale behind combining these two methods is to refine the set of rules produced by association rule learning. That is, each cluster has different patterns of relationships to other clusters. Since the generated rules are ordered by relevance (confidence) this gives explicit support to the pertinence of the attributes identified in the related rules. The extracted pertinent features will be considered in the Bio-PEPA model refinement by both expert and developer. For instance, if the attribute Sex is identified as pertinent then the initial Bio-PEPA model is enhanced by including Sex in the model, where compartments are restructured, parameters recalculated and functional rates redefined.

### *3.4.1 Application to Mumps: clustering for observed data features*

Fig. 2 shows the clusters formed by the K-means algorithm on the data OneYearOneBoardVaccStatus using a minimum cluster size of two and without any predefined class attribute, as our data is only suitable for unsupervised learning. Fig. 2 shows a plot of the attribute Week versus the attribute MMR status (see Appendix A for additional plots of attributes). It clearly illustrates that the set of instances are assigned to two disjoint spaces related to the

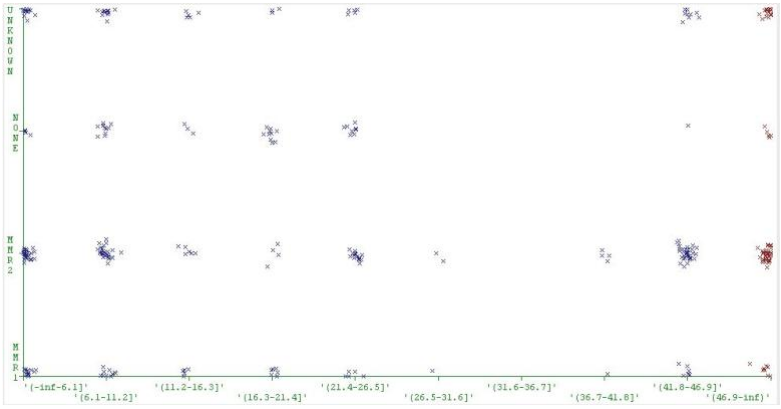


**Fig. 2** K-means clustering applied to mumps confirmed cases OneYearOneBoardVaccStatus. There are two clusters identified (blue and black crosses)

attribute week. Most of the instances are clustered from week 1 to week 26 and from week 41 to the last week of the year. The majority of instances are deemed to be in the line of the attribute value MMR status = MMR2. MMR2 denotes fully vaccinated status, i.e., two doses of MMR vaccine most likely given at 13 months and 3 years of age. Note that the weeks have been automatically discretized into 10 groups. No instances have been detected as outliers. Increasing the cluster number to three produces no new insight (same results as Fig. 2).

### *3.4.2 Application to Mumps: association rules for observed data features*

Once clusters have been identified, association rules are generated per cluster. Applying clustering as a prior step to association rule learning leads to increased confidence in the resulting rules. Association rules learning is applied to OneYearOneBoardVaccStatus (clusters as above) using Apriori (resulting in five rules), Predictive Apriori (18 rules) and Tertius (11 rules) algorithms. MMR status and its relation with Week and Age is important in 29 of the 34 rules produced: this gives confidence in the results. Selected results are shown in Table 2. Where Predictive Apriori and Tertius give the same results as Apriori these rules are omitted. Similarly, we omit rules concerning Sex as it is distributed across cases and is not implicated in mumps acquisition. For example, in Table 2, rules 5 and 7 are redundant in the presence of rule 6. Note the high confidence levels of these results. Indeed, repeating the experiment in WEKA without prior clustering will produce lower confidence (80%) for the same resulting rules. These generated rules show clearly the relationship between contracting mumps, having had the MMR vaccine (MMR1 or MMR2), and being in the age group 16-25. Combined with Rule 4 this suggests a link between college and university terms in Scotland, and disease outbreak.

**Table 2** Association Rule results on clustered data OneYearOneBoardVaccStatus

| **Algorithm: Apriori** Rules | Measure |
|---|---|
| 1. Week=‘(41.8-46.9]’ and MMR STATUS=MMR2 ⇒ Age=‘(16.8-25.2]’ | conf:(1) |
| 2. Week=‘(46.9-inf)’ and MMR STATUS=MMR2 ⇒ Age=‘(16.8-25.2]’ | conf:(0.97) |
| 3. Week=‘(6.1-11.2]’ and MMR STATUS=MMR2 ⇒ Age=‘(16.8-25.2]’ | conf:(0.96) |
| 4. Week=‘(41.8-46.9]’ ⇒ Age=‘(16.8-25.2]’ | conf:(0.96) |
| 5. Sex=F and MMR STATUS=MMR2 ⇒ Age=‘(16.8-25.2]’ | conf:(0.95) |
| 6. MMR STATUS=MMR2 ⇒ Age=‘(16.8-25.2]’ | conf:(0.95) |
| 7. SEX=M and MMR STATUS=MMR2 ⇒ Age=‘(16.8-25.2]’ | conf:(0.95) |
| **Algorithm: Predictive Apriori** Rules | Measure |
| 1. Week=‘(46.9-inf)’ and MMR STATUS=MMR1 ⇒ Age=‘(16.8-25.2]’ | acc:(0.98136) |
| 2. Sex=F and Week=‘(16.3-21.4]’ and MMR STATUS=MMR1 ⇒ Age=‘(16.8-25.2]’ | acc:(0.90842) |
| 3. Week=‘(-inf-6.1]’ and MMR STATUS=MMR1 ⇒ Age=‘(16.8-25.2]’ | acc:(0.90771) |

The rules and clusters identified strongly suggest that seasonal variation of transmission should be included in the SEIR Bio-PEPA model, and that two new classes of individuals should be included who have been vaccinated either once or twice but who still get mumps (therefore their immunity must wane). This is illustrated in Bio-PEPA in the next section. Age structure could be added as this is also relevant; however, for our data there is a strong correlation between age and MMR status (due to the time at which vaccination began) therefore we choose to include only vaccination status at this point.

### 3.5 Mumps Bio-PEPA model optimisation

Taking the initial SEIR compartmental model [2], the Bio-PEPA model is refined manually to include seasonality and vaccination as suggested by clustering and association rule mining. Fig. 3 shows the model extended with new compartments. In total, seven compartments describe the model: vaccinated individual (V1 related to the 1st dose of MMR and V2 related to the 2nd dose of MMR), susceptible (susceptible 1 related to native susceptible and susceptible 2 related to modified susceptible due to waning vaccine immunity), exposed (individual who acquired the virus from either susceptible 1 or 2 but is not yet infectious), infected (infectious individual) and recovered individual. We consider a homogeneous well-mixed population. Infection can arise locally or through immigration. The parameters depicted in Fig. 3 as transition labels and described in Table 3 define the rate of transition between each compartment. According to the Bio-PEPA formalism, the mumps model as shown in Fig. 3 is expressed by 12 parameters (an additional transmission rate $\beta_3$ replacing $\beta_2$ and $\beta_1$ according to season). The proposed parameter values are

**Table 3** Mumps model parameters

| Parameter | Description | Value (day) | Formula |
|---|---|---|---|
| B | Birth rate | $3\ 10^{-5}$ | Number of birth / Total population |
| $\mu$ | Death rate | $3.7\ 10^{-5}$ | Number of death / Total population |
| $\mu_1$ | Non-vaccination rate | $2.1\ 10^{-6}$ | Birth rate Birth rate $-(\mu_2 + \mu_3)$ |
| $\mu_2$ | Vaccination rate (MMR1) | $2.8\ 10^{-6}$ | Birth rate * MMR1 vaccination coverage |
| $\mu_3$ | Vaccination rate (MMR2) | $2.5\ 10^{-5}$ | Birth rate * MMR2 vaccination coverage |
| $\tau$ | Waning immunity rate (MMR1) | $3.4\ 10^{-4}$ | 1/immunity duration of MMR1 |
| $\delta$ | Waning immunity rate (MMR2) | $\tau/2$ | 1/immunity duration of MMR2 |
| | Transmission rates : | | transmission rate $= R_0 * 1/\gamma$ |
| $\beta_1$ | - High season and native susceptible | $C1 * \beta_2$ | where $R_0$ is the basic reproductive rate, $R_0 \in [4, 11]$ |
| $\beta_2$ | - High season and modified susceptible | $R_0 * 1/\gamma$ | C1 and C2 are constant scaling factors (C1=0.78, C2=0.44) obtained experimentally |
| $\beta_3$ | - Low season | $C2 * \beta_2$ | |
| $\alpha$ | Incubation period | $[12 - 25]$ | 1/infection rate |
| $\gamma$ | Infection period | $[7 - 9]$ | 1/recovery rate |
| $\lambda$ | Immigration rate | 0.07 | Immigration$*\sqrt{population}$[46] |

summarized from the literature [2,45,46] or deduced from serological study by epidemiologists [47]. For modelling convenience, we assume the same period of infection and incubation [47] for both natively susceptible and modified susceptible. The Bio-PEPA description is given in Appendix B.

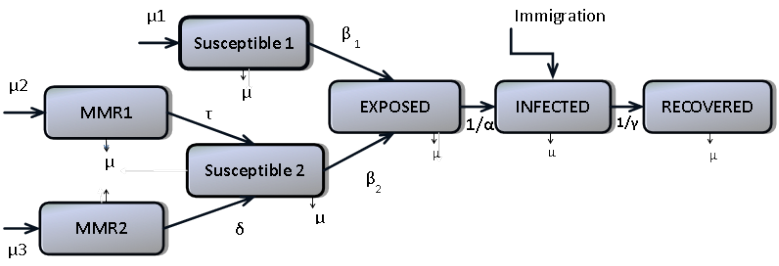


**Fig. 3** Mumps compartmental model

### 3.6 Data mining for optimal parameters

Our aim here is to identify Bio-PEPA model outputs (simulations results related to the daily predicted cases) giving a close match to observed data (daily reported cases). Depending on the similarity measure, the best cluster would include two objects: one model output relating to the optimal parameter set, and the observed data. Thus, overlapped and probabilistic techniques are of no interest here as objects can be assigned to several clusters.

Algorithm 1 describes how the Bio-PEPA model and experimental simulation output is combined with clustering. Essentially, a broad parameter sweep is carried out, generating a collection of experimental results as time series of daily predicted cases. Then, clustering is used to group these and the observed data. The goal of the exercise is to determine patterns in parameter settings related to matching time series outputs. Bio-PEPA provides two simulation techniques providing time series output: deterministic simulation and stochastic simulation (where an average of multiple runs is considered). Either could be used in our approach.

As shown in algorithm 1, we define the output of Bio-PEPA as the input of clustering, where each simulation experiment $Exp_i$ corresponds to a set of parameters $G_i$ and an instance/line (individual) in the clustering table. In clustering each instance is defined by a set of attributes (columns). As time series data are a function of simulation times (from 0 to $T$ - the end of simulation), the attributes reflect each data point. Therefore, the input of the clustering algorithm will be a table $B[j,t]$, where $j$ is the number of simulation experiments plus the observed data and $t$ is the simulation time. A pre-condition for this algorithm is that the simulation experiments should be structured in the same way as the observed data (where the number of infected are daily reported cases).

Clustering is applied by aiming for the smallest number of groups first, and refining until the cluster $C_m$ (including the observed data) has at most one other member. If the algorithm terminates with $C_m$ containing only the observed data then the previous value of $C_m$ should be returned. All members of $C_m$ are in some way optimal parameter settings. As each experiment is related to a set of model parameters then the other clusters generated in this process may be useful to give contextual information about patterns relating to input parameters of the model. For example, a cluster may identify a range of parameter settings which all give similar outputs, therefore the model is not sensitive to that parameter.

#### *3.6.1 Application to mumps: clustering for parameter optimisation*

Before suggesting any future trends to the experts, we should first convince them of the validity of our model. Thus, we focus first on reproducing the dynamics of mumps from 2004 to 2015 in Scotland.

We carried out a series of simulations in the Bio-PEPA plug-in. Deterministic simulation is used to efficiently provide a simple, consistent, comparable

Inputs

1. Select Bio-PEPA model parameters to be investigated and range of values. N = number of parameters * number of values;
2. Set the group $G_i$ to the $i$th series of parameter values used to run the $i$th Bio-PEPA experiment, where $i \in [1, N]$;
3. Set the experiment $Exp_i$ to the simulation results relating to the group $G_i$, and let $T$ be the time of simulation end, where each $Exp_i[t]|t \in [0, T]$ corresponds to a time point;
4. Let $Exp_{N+1}$ be the time series of the observed data;
5. Let $B$ be the constructed database, where each row $B[j] = Exp_j$ is the $j : j^{th}$ experiment for $j \in [1, N + 1]$;

Outputs

The cluster $C_m$ containing observed data and fitting experiment.

Algorithm

6. Define target number of clusters $K \in [2, N - 1]$;
7. Initialise $K = 2$;
8. Apply the clustering algorithms using $B$ as an input. The output is a set $\{C_k|k \in [1, K]\}$ of clusters. Identify $C_m$ as the cluster containing $Exp_{N+1}$;
9. If $|C_m| > 2$ then increment $K$ and repeat from step 8 until $|C_m| \leq 2$, or no more clustering occurs.

**Algorithm 1:** Optimisation of parameters of a Bio-PEPA model

time series output. Stochastic simulation (unless averaged over multiple runs) will provide varying results for a single parameter set. The time required for multiple runs makes stochastic simulation unsuitable for the current example.

The simulation is of 12 years, starting at t=0 (related to the start of the year 2004) and ending at t=4320 (related to the end of the year 2015, assuming months have 30 days for simplicity and therefore years have 360 days). The time unit is days. To select the optimal simulation result, we first identify relevant parameters and their range of values. The parameters of interest are: incubation period $\alpha$, infectious period $\gamma$, and basic reproductive rate $R_0$. As the transmission rates $\beta$ are defined by $(R_0 * 1/\gamma)$, their values vary as a function of infectious period and $R_0$. See Table 4 for parameter ranges. Both $\alpha$ and $\gamma$ are measured in days therefore increments of one are logical. $R_0$ might reasonably be measured with more granularity; however, a variation of one step for each parameter results in 14 points for incubation period, 4 points for infectious period and 8 points for $R_0$. This would give 448 (14*4*8) simulation experiments. Each run is short, but the translation to WEKA is time-consuming. Assuming model behaviour is continuous, we identify a sample across the largest range to reduce the number of experiments required initially, allowing promising parameter regions to be identified for further investigation. Table 4 specifies the ranges used. This results in 160 (5*4*8) experiments. A total of 699040 (160*4321) time points are to be considered in clustering the model simulation output with the observed data.

Table 5 summarises the results of the clustering algorithms. All algorithms except sIB eventually returned the target of a cluster of two where one object is the observed data. Moreover, the same experiment was returned in this

**Table 4** Experimental range of values for parameters

| Parameter | Initial range | Experimental sample | Optimal value |
|---|---|---|---|
| $\alpha$ | 12-25 | 12, 15, 18, 21, 25 | 21 |
| $\gamma$ | 6-9 | 6-9 | 7 |
| $R_0$ | 4-11 | 4-11 | 6 |

**Table 5** Clustering analysis results

| Clustering Algorithm | Number of clusters (K) | Time processing (seconds) |
|---|---|---|
| K-means | 36 | 4.06 |
| Hierarchical clustering | 5 | 2.03 |
| FarthestFirst | 7 | 0.22 |
| EM clustering | 15 | 45.91 |
| X-means | 15 | 11.28 |

cluster for every successful algorithm (with the optimal values as in column 4 of Table 4). The Hierarchical and FarthestFirst algorithms returned our target within a small number of clusters (7 or fewer). K-means, the most popular clustering algorithm, took longer: the desired cluster only appeared when the target was 36 clusters. All algorithms run in under one minute for the largest number of clusters for this data. In contrast, sIB takes significantly longer (around nine minutes) and the smallest cluster containing the observed data has two other experiments. However, one of the included experiments found by the sIB algorithm has similar parameter values to those noted by other algorithms. All algorithms use Euclidean distance as a similarity measure, except sIB, which uses Kullback-Leibler divergence.

**Table 6** K-means clustering: parameter values associated with 10 year cycles

| K-MEANS CLUSTERING | | | | | |
|---|---|---|---|---|---|
| No Cluster | No object in cluster | Cycle | Parameter values | | |
| | | | $R_0$ | $\alpha$ | $\gamma$ |
| 1 | 2 | 9,10 | 7 | 21,25 | 9 |
| 5 | 8 | 8-10 | 6,7 | 12-21 | 7-9 |
| 6 | 3 | 9-10 | 7-8 | 25 | 7-9 |
| 10 | 7 | 9-10 | 5-7 | 12-25 | 6-7 |
| 17 | 2 | 10 | 6 | 21 | 7 |
| 23 | 3 | 9-11 | 5 | 12-15 | 8-9 |
| 26 | 3 | 10-11 | 6 | 21-25 | 8-9 |

Considering clusters not containing the observed data, we see significant differences in the length of the major cycle depending on the value of $R_0$. (Recall that the observed data of Fig. 1 shows epidemic outbreaks every year and a major cycle of 10 years.) For example the hierarchical clustering algorithm defined four other clusters which show cycles of 6, 7, 9 and 12 years. In fact, the period between major outbreaks varies inversely with $R_0$. Table 6 presents

a selected sample of clusters from the K-means algorithm showing how the parameters vary even for cycles of period 9-11. Higher values of $\alpha$ contribute, with $R_0$, to longer inter-epidemic periods (see e.g. lines 17 and 26). There appears to be no effect from varying $\gamma$.

Recall that the values of $\alpha$ were sampled across the parameter range. We repeat the clustering exercise with more granularity. Fig. 4 plots time series simulation results for $\alpha \in [19, 24]$. Clustering confirms that 21 is the optimal value for $\alpha$ by grouping that experiment with the observed data. The optimal parameter set is shown in column 4 of Table 4. Intuitively, a long incubation period seems reasonable, as the number of individuals in the exposed compartment accumulates individuals and leads to a large reservoir for new cases.

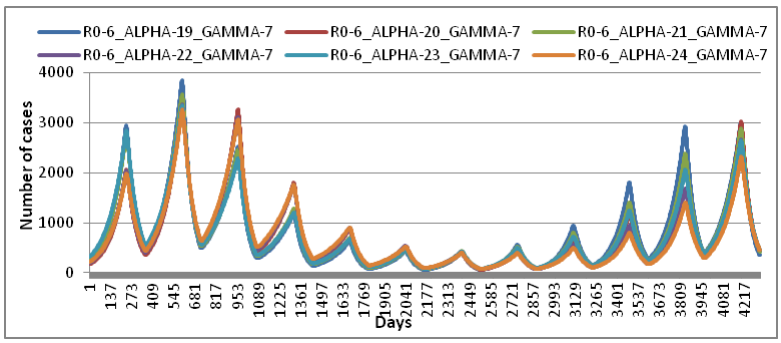


**Fig. 4** Simulation traces for $\alpha \in [19, 24]$

Finally, we also examined clusters obtained in the penultimate step of clustering containing the observed and more than one object (experiment) given by both K-means and sIB algorithms. Figs. 5 and 6 illustrate that these time series cannot be easily discriminated by eye: the basic pattern of 10 year cycles is maintained, but there are small differences in amplitude. For example, Fig. 6 depicts three graphs related to the three objects belonging to the same cluster in the step prior to providing the targeted results.

Fig. 7 plots observed data and time series simulation results for the optimal set, where the simulation results are scaled to account for under-reporting of disease [48]. Scaling makes no difference to the clustering analysis: it was performed with and without scaling data, and similar results were obtained. The aim is to match the general shape of outbreaks, which this does well apart from the years 2009 and 2010. A larger $R_0$ would result in shorter cycles (as illustrated above), but while this might match 2009, it wouldn't match both 2009 and 2010. Clustering, with its underlying statistical comparison measures, has automatically highlighted the major periodic cycle of ten years. Running the model for a longer period (not shown here) confirms ten year cycles. This suggests a resolution to our initial question about mumps concerning the future trend of outbreaks. Paired with the use of association rules and clustering

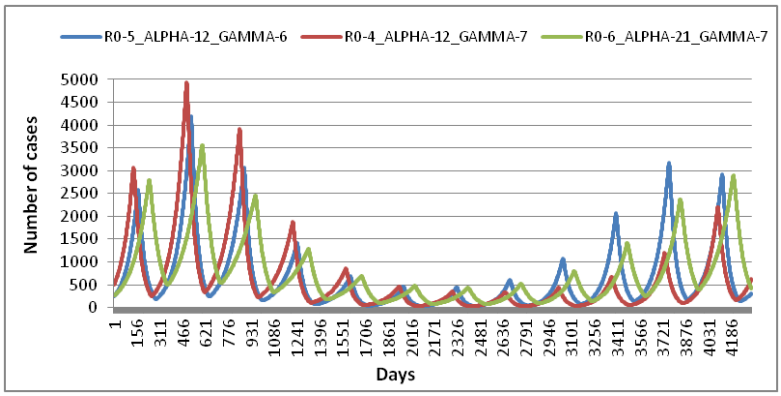


**Fig. 5** Comparison of experiments assigned to the same cluster by sIB algorithm (step prior to the one providing the targeted results)

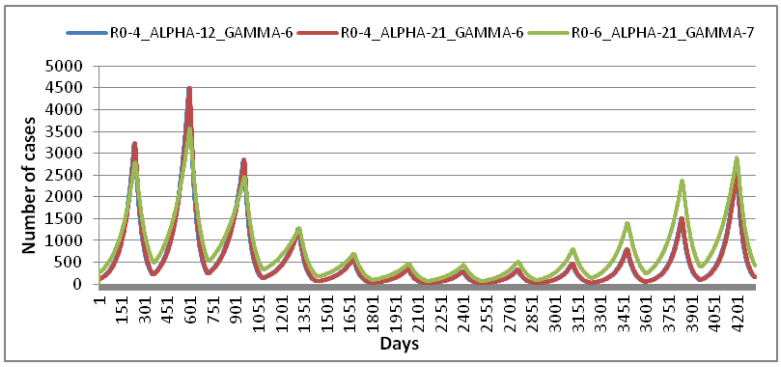


**Fig. 6** Comparison of experiments assigned to the same cluster by K-means algorithm (step prior to the one providing the targeted results)

to identify and explain model structure this leads to a model which epidemiologists and experts can be comfortable in using for future predictions and decision-making.

## 4 Discussion and Conclusion

In this paper we have presented results demonstrating the utility of combining Bio-PEPA modelling with data mining applied to mumps in both model structure identification and model parameter optimization. By innovatively preceding association rule learning by clustering we were able to identify strong features in the data suggesting areas in which a standard model could be refined. Specifically, this process suggested that seasonality, vaccination and age are correlated features which should be included in the model. We carried out a series of simulations from the manually crafted model to predict outbreaks

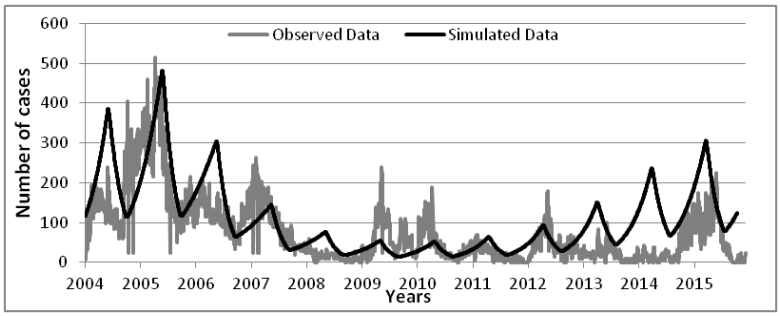


**Fig. 7** Predicted incidence of Mumps in Scotland from 2004 to 2015 (Bio-PEPA simulation results related to the selected experiment – $\alpha = 21$,$\gamma = 7$, $R_0 = 6$) and observed data. The simulated data are scaled to fit the observed data.

from 2004 to 2015 in Scotland. From 164 time series, where each one is related to specific parameter values, our proposed algorithm could identify the one best fitting to observed data. Conversely, although many of the remaining experiments seem to match well by eye with observed data, they have not been selected by the algorithms as well-fitted patterns to observed data. Thus, analysing Bio-PEPA outputs using clustering reduces the uncertainty in model parameter estimation. Several clustering algorithms were used. The fact that the majority return the same result gives more confidence that this is the optimal parameter set within the proposed series of experiments. Moreover, the clusters produced can be examined to detect trends in parameter values. By using Bio-PEPA modelling and its simulation tool combined with clustering techniques, we were able, not only to add confidence to our model construction and prediction, but also to shed light on the selected values which can help epidemiologists to understand the dynamic of mumps epidemics and then make better decisions.

This work can be extended in several directions. The basic data mining algorithms built in WEKA, with default parameters, were used. The mumps example could be re-analysed with some of the specialised time series clustering algorithms, although, on the whole, these are about speeding-up clustering and dealing with warp between time series. Within epidemiology, a more fine-grained approach to parameter selection is desirable: further work can establish how far clustering can help when the difference between parameter values is very small (e.g. increments of 0.1, or smaller). For this particular example, we might also consider allowing the constants C1 and C2 to vary, increasing or decreasing seasonal effects. Further developments would require epidemiologists to collect more detailed data. For example, the model could be enhanced with population movements and wider immigration and emigration, but this would require additional spatial population information. Lastly, application to other data sets would be informative. We propose here epidemiological models of mumps but there is no reason that this approach could not be extended to

models of other types of infectious disease, or systems producing time series outputs more generally.

## 5 Acknowledgments

We thank the anonymous reviewers for critical support and review of the manuscript.

## References

1. T. R. Malthus, *An essay on the principle of population: or, A view of its past and present effects on human happiness*. Reeves & Turner, 1888.
2. R. M. Anderson and R. M. May, *Infectious diseases of humans: dynamics and control*, vol. 28. Wiley Online Library, 1992.
3. S. Abrams, P. Beutels, and N. Hens, "Assessing mumps outbreak risk in highly vaccinated populations using spatial seroprevalence data," *American journal of epidemiology*, pp. 1006–17, 2014.
4. H. W. Hethcote, "The mathematics of infectious diseases," *SIAM review*, vol. 42, no. 4, pp. 599–653, 2000.
5. C. Castillo-Chavez, S. Blower, P. Driessche, D. Kirschner, and A.-A. Yakubu, *Mathematical approaches for emerging and reemerging infectious diseases: models, methods, and theory*. Springer, 2002.
6. A. Vespignani, "Modelling dynamical processes in complex socio-technical systems," *Nature Physics*, vol. 8, no. 1, pp. 32–39, 2012.
7. A. C. Babtie, P. Kirk, and M. P. Stumpf, "Topological sensitivity analysis for systems biology," *Proceedings of the National Academy of Sciences*, vol. 111, no. 52, pp. 18507–18512, 2014.
8. T. Asha, S. Natarajan, and K. Murthy, "Data mining techniques in the diagnosis of tuberculosis," in *Understanding Tuberculosis-Global Experiences and Innovative Approaches to the Diagnosis*, InTech, 2012.
9. H. H. Inbarani, A. T. Azar, and G. Jothi, "Supervised hybrid feature selection based on pso and rough sets for medical diagnosis," *Computer methods and programs in biomedicine*, vol. 113, no. 1, pp. 175–185, 2014.
10. H. Kim, M. I. M. Ishag, M. Piao, T. Kwon, and K. H. Ryu, "A data mining approach for cardiovascular disease diagnosis using heart rate variability and images of carotid arteries," *Symmetry*, vol. 8, no. 6, p. 47, 2016.
11. F. Ciocchetta and J. Hillston, "Bio-PEPA: A framework for the modelling and analysis of biological systems," *Theoretical Computer Science*, vol. 410, no. 33-34, pp. 3065–3084, 2009.
12. E. Bartocci and P. Lió, "Computational modeling, formal analysis, and tools for systems biology," *PLoS Comput Biol*, vol. 12, no. 1, p. e1004591, 2016.
13. M. L. Guerriero, "Qualitative and quantitative analysis of a Bio-PEPA model of the gp130/jak/stat signalling pathway," in *Transactions on Computational Systems Biology XI*, pp. 90–115, Springer, 2009.
14. D. Hamami, A. Baghdad, and C. Shankland, "Decision support based on Bio-PEPA modeling and decision tree induction: a new approach, applied to a tuberculosis case study," *International Journal of Information Systems in the Service Sector (IJISSS)*, vol. 9, no. 2, pp. 71–101, 2017.
15. D. Hamami and B. Atmani, "Obtaining optimal Bio-PEPA model using association rules: Approach applied to tuberculosis case study," in *International Conference on Information Systems for Crisis Response and Management in Mediterranean Countries*, pp. 62–75, Springer, 2016.
16. P. M. Pardalos, V. L. Boginski, and V. Alkis, *Data mining in biomedicine*, vol. 7. Springer Science & Business Media, 2008.

17. M. Sebban, I. Mokrousov, N. Rastogi, and C. Sola, "A data-mining approach to spacer oligonucleotide typing of mycobacterium tuberculosis," *Bioinformatics*, vol. 18, no. 2, pp. 235–243, 2002.
18. S. M. Lynch and J. H. Moore, "A call for biological data mining approaches in epidemiology," *BioData mining*, vol. 9, no. 1, p. 1, 2016.
19. B. J. Ross and J. Imada, "Evolving stochastic processes using feature tests and genetic programming," in *Proceedings of the 11th Annual conference on Genetic and evolutionary computation*, pp. 1059–1066, ACM, 2009.
20. D. R. Oaken, *Optimisation of Definition Structures Parameter Values in Process Algebra Models Using Evolutionary Computation*. PhD thesis, University of Stirling, 2014.
21. E. Bartocci, L. Bortolussi, and G. Sanguinetti, "Data-driven statistical learning of temporal logic properties," in *International Conference on Formal Modeling and Analysis of Timed Systems*, pp. 23–37, Springer, 2014.
22. T. Sumner, *Sensitivity analysis in systems biology modelling and its application to a multi-scale model of blood glucose homeostasis*. PhD thesis, UCL (University College London), 2010.
23. C. Okaïs, S. Roche, M.-L. Kürzinger, B. Riche, H. Bricout, T. Derrough, F. Simondon, and R. Ecochard, "Methodology of the sensitivity analysis used for modeling an infectious disease," *Vaccine*, vol. 28, no. 51, pp. 8132–8140, 2010.
24. J. L. Moore, S. Liang, A. Akullian, and J. V. Remais, "Cautioning the use of degree-day models for climate change projections in the presence of parametric uncertainty," *Ecological Applications*, vol. 22, no. 8, pp. 2237–2247, 2012.
25. R. Hickson, G. Mercer, K. Lokuge, *et al.*, "Sensitivity analysis of a model for tuberculosis," in *19th international congress on modelling and simulation*, pp. 926–932, 2011.
26. J. Wu, R. Dhingra, M. Gambhir, and J. V. Remais, "Sensitivity analysis of infectious disease models: methods, advances and their application," *Journal of The Royal Society Interface*, vol. 10, no. 86, p. 20121018, 2013.
27. A. Georgoulas, J. Hillston, D. Milios, and G. Sanguinetti, "Probabilistic programming process algebra," in *International Conference on Quantitative Evaluation of Systems*, pp. 249–264, Springer, 2014.
28. L. Bortolussi, D. Milios, and G. Sanguinetti, "Smoothed model checking for uncertain continuous-time markov chains," *Information and Computation*, vol. 247, pp. 235–253, 2016.
29. E. Bartocci, L. Bortolussi, L. Nenzi, and G. Sanguinetti, "System design of stochastic models using robustness of temporal properties," *Theoretical Computer Science*, vol. 587, pp. 3–25, 2015.
30. T. W. Liao, "Clustering of time series data — a survey," *Pattern recognition*, vol. 38, no. 11, pp. 1857–1874, 2005.
31. F. Ciocchetta and J. Hillston, "Bio-PEPA for epidemiological models," *Electronic Notes in Theoretical Computer Science*, vol. 261, pp. 43–69, 2010.
32. S. Benkirane, R. Norman, E. Scott, and C. Shankland, "Measles epidemics and PEPA: an exploration of historic disease dynamics using process algebra," in *International Symposium on Formal Methods*, pp. 101–115, Springer, 2012.
33. D. Hamami and B. Atmani, "modeling the effect of vaccination on varicella using Bio-PEPA," in *International Conference on Modeling and Simulation MS2012*, pp. 783–077, Proc IASTED, 2012.
34. D. Hamami and B. Atmani, "Tuberculosis modelling using Bio-PEPA approach," *World Academy of Science, Engineering and Technology, International Journal of Medical, Health, Biomedical, Bioengineering and Pharmaceutical Engineering*, vol. 7, no. 4, pp. 183–190, 2013.
35. A. Ramanathan, C. A. Steed, and L. L. Pullum, "Verification of compartmental epidemiological models using metamorphic testing, model checking and visual analytics," in *BioMedical Computing (BioMedCom), 2012 ASE/IEEE International Conference on*, pp. 68–73, IEEE, 2012.
36. I. Bonmarin, P. Santa-Olalla, and D. Lévy-Bruhl, "Modélisation de l'impact de la vaccination sur l'épidémiologie de la varicelle et du zona," *Revue d'epidemiologie et de sante publique*, vol. 56, no. 5, pp. 323–331, 2008.

37. A. L. De Espíndola, C. T. Bauch, B. C. T. Cabella, and A. S. Martinez, "An agent-based computational model of the spread of tuberculosis," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2011, no. 05, p. P05003, 2011.
38. R. Sullivan, *Introduction to data mining for the life sciences*. Springer Science & Business Media, 2012.
39. I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.
40. D. Tomar and S. Agarwal, "A survey on data mining approaches for healthcare," *International Journal of Bio-Science and Bio-Technology*, vol. 5, no. 5, pp. 241–266, 2013.
41. M. Donaghy, J. C. Cameron, and V. Friederichs, "Increasing incidence of mumps in scotland: options for reducing transmission," *Journal of clinical virology*, vol. 35, no. 2, pp. 121–129, 2006.
42. R. L. Cameron and A. Smith-Palmer, "Measles, mumps, rubella and whooping cough illness, routine childhood vaccine uptake," tech. rep., Health Protection Scotland, 2015.
43. R. L. Cameron and A. Smith-Palmer, "Measles, mumps, rubella and whooping cough illness, routine childhood vaccine uptake," Tech. Rep. 01, Health Protection Scotland, 2016.
44. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.
45. M. J. Keeling and P. Rohani, *Modeling infectious diseases in humans and animals*. Princeton University Press, 2008.
46. B. Finkenstädt, M. Keeling, and B. Grenfell, "Patterns of density dependence in measles dynamics," *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 265, no. 1398, pp. 753–762, 1998.
47. P. Morgan-Capner, J. Wright, C. L. Miller, and E. Miller, "Surveillance of antibody to measles, mumps, and rubella by age.," *BMJ*, vol. 297, no. 6651, pp. 770–772, 1988.
48. A. Takla, O. Wichmann, C. Klinc, H. Hautmann, T. Rieck, and J. Koch, "Mumps epidemiology in germany 2007-11," *Eurosurveillance*, vol. 18, no. 33, p. 20557, 2013.

## A Additional figures

Fig. 8 (resp. Fig. 9 and Fig. 10) shows that K-means clustering applied to mumps confirmed cases and plotted by Age (resp. MMR status and Sex) is not meaningful for the model where both clusters depicted similar age groups (resp. similar distribution of MMR status and Sex).

## B Model

1. mu = 0.000037;
2. beta1 =0.7;
3. beta2 =0.9;
4. beta3 = 0.4;
5. mu1 = 0.0000021;
6. mu2= 0.0000028;
7. mu3= 0.000025;
8. alpha = 0.05;
9. gamma = 0.143;
10. lambda =0.07;
11. tau= 0.00034;
12. delta=tau/2;
13. sizeOutside = 110000;
14. sizeLocal = 5300000;
15. location world : size =5200000 , type = compartment;
16. location Local in world: size = sizeLocal, type = compartment;

17. location Local in world: size = sizeLocal, type = compartment;
18. location Outside in world : size = sizeOutside, type = compartment;
19. thigh = 4;
20. tlow = 9;
21. month = floor(time/30);
22. season_time = 1-H( ((month - 12*floor(month/12)) - tlow)* (thigh-(month - 12*floor(month/12)))
    );
23. N = (S1@Local +E@Local + I@Local + R@Local +S2@Local + MMR1@Local +
    MMR2@Local);

    Kinetic Laws
24. kineticLawOf BIRTH1: mu1 * N;
25. kineticLawOf BIRTH2: mu2 * N;
26. kineticLawOf BIRTH3: mu3 * N;
27. kineticLawOf MMR1_S2: MMR1@Local *tau;
28. kineticLawOf MMR2_S2: MMR2@Local *delta;
29. kineticLawOf Death_MMR1 : mu * MMR1@Local;
30. kineticLawOf Death_MMR2 : mu * MMR2@Local;
31. kineticLawOf immigration : lambda/10000;
32. kineticLawOf S1_E: (beta1 * S1@Local * I@Local)/N * (season_time) + (1-season_time)*(beta3
    * S1@Local * I@Local)/N ;
33. kineticLawOf S2_E: (beta2 * S2@Local * I@Local)/N * (season_time) + (1-season_time)*
    (beta3 * S2@Local * I@Local)/N;
34. kineticLawOf E_I: alpha * E@Local;
35. kineticLawOf I_R: gamma * I@Local;
36. kineticLawOf Death_S1: mu * S1@Local;
37. kineticLawOf Death_I: mu * I@Local ;
38. kineticLawOf Death_E: mu * E@Local;
39. kineticLawOf Death_S2: mu * S2@Local;
40. kineticLawOf Death_R: mu * R@Local;

    Species
41. S1 = (BIRTH1,1) >> S1@Local + (S1_E,1) << S1@Local + Death_S1 << S1@Local;
42. S2 = (S2_E,1) << S2@Local + Death_S2 << S2@Local + (MMR2_S2,1) >> S2@Local
    +(MMR1_S2,1) >> S2@Local;
43. E = (S1_E,1) >> E@Local +(S2_E,1) >> E@Local +(E_I,1) << E@Local
    + Death_E << E@Local;
44. I = (E_I,1) >> I@Local +(I_R,1) << I@Local + Death_I << I@Local
    + immigration[Outside → Local](.)I + (S1_E,1) (.) I+ (S2_E,1) (.) I;
45. R = (I_R,1) >> R@Local+ Death_R << R@Local ;
46. MMR1 = (BIRTH2,1) >> MMR1@Local + (MMR1_S2,1) << MMR1@Local
    + Death_MMR1 << ;
47. MMR2 = (BIRTH3,1)>> MMR2@Local + (MMR2_S2,1) << MMR2@Local
    + Death_MMR2 << ;

    Model component
48. S1@Local[1100000]< ∗ > S2@Local[0]< ∗ > E@Local[0] < ∗ > I@Local[20]
    < ∗ > R@Local[3218600] < ∗ > MMR1@Local[273541] < ∗ > MMR2@Local[250000]
    < ∗ > I@Outside[10000]

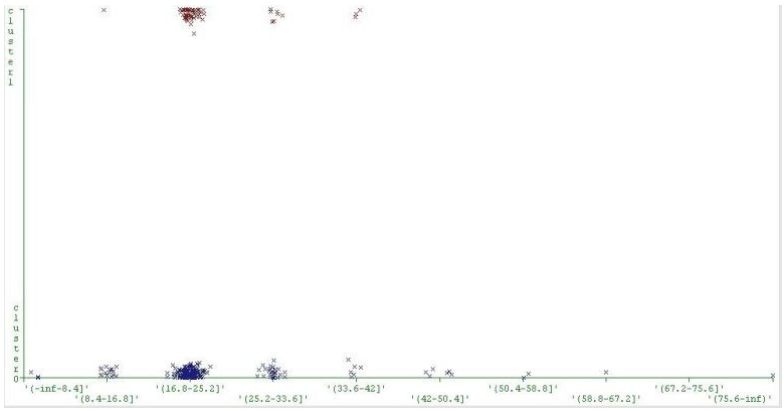


**Fig. 8** K-means clustering applied to mumps confirmed cases OneYearOneBoardVaccStatus: Clusters vs Age.

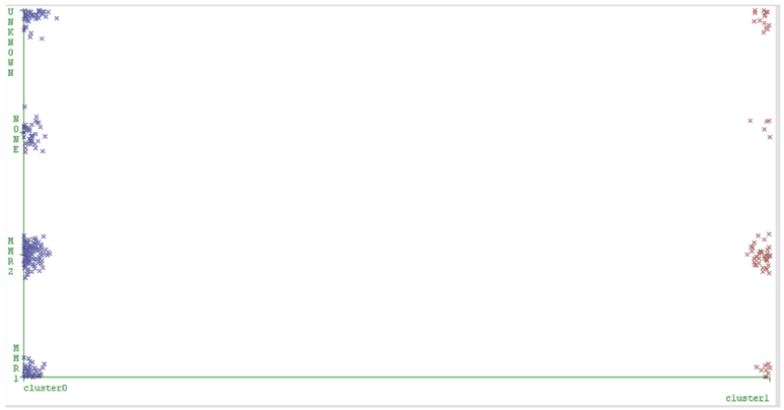


**Fig. 9** K-means clustering applied to mumps confirmed cases OneYearOneBoardVaccStatus: Clusters vs MMR Status.

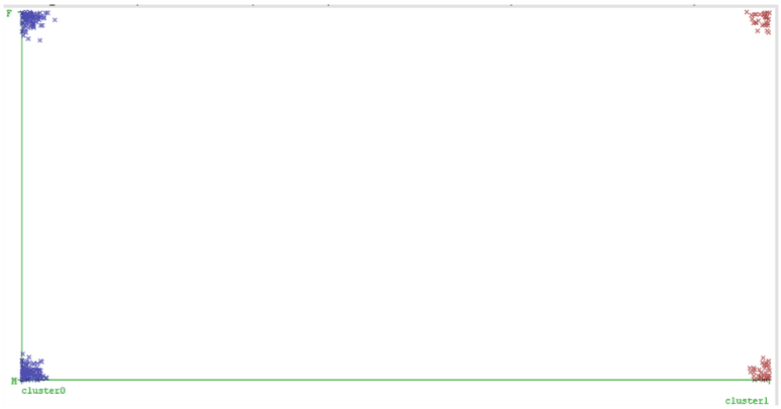


**Fig. 10** K-means clustering applied to mumps confirmed cases OneYearOneBoardVaccStatus: Clusters vs Sex.