Antibiotics Resistance Forecasting: A Comparison of Two Time Series Forecast Models

Darja Strahlberg
Hamburg University of Technology, Germany
darja.strahlberg@gmail.com

Advisors: Prof. Dr. Michael Kolbe, Center for Structural Systems Biology,
University of Hamburg, Germany
michael.kolbe@cssb-hamburg.de
Gianni Pagnini, Basque Center for Applied Mathematics & Ikerbasque,
Bilbao, Basque Country - Spain
gpagnini@bcamath.org

August 21, 2021

Abstract

The rise of antibiotic resistance is a growing challenge for global health. Antibiotics are used for disease treatment, as well as for medical procedures, for instance, operations and transplants. The aim of this work is to compare auto-regressive integrated moving average (ARIMA) and recurrent neural networks (RNN) to forecast the spread of drug-resistant bacterial infections at the community level. The comparison of two algorithms is performed for a multi-step time series univariate dataset. Five distinct time series were modelled, each one representing the number of episodes per single ESKAPE infecting pathogen, that has occurred quarterly between 2008 and 2018 calendar years in Germany. The forecast quality is evaluated by the root mean squared error between the forecasted values and the test data set. The experimental results show that multi-neural network forecasting RNN is significantly poorer than ARIMA for multi-step forecasting on univariate datasets. Finally, the paper provides a conclusion, that machine learning complexity is not always adding skill to the forecast. The forthcoming challenges are setting conditions when machine learning models can perform well for the real-world applications. The code used to evaluate the concept is available.

1 Introduction

The discovery of antibiotics is considered one of the most significant health-related events of the last century. Since the introduction of penicillin, the deployment of any novel antibiotic has been followed by the evolution of clinically relevant resistance strains to that antibiotic in as little as a few years [2]. In addition to their use in the treatment of infectious diseases, antibiotics are critical for the success of advanced surgical procedures, including organ and prosthetic transplants [4]. So emerging antibiotic resistance is a much broader problem than initially anticipated.

Forecasting the spread of drug-resistant bacterial infections helps in developing context-specific national and regional operational plans. The accurate forecasting of future behaviour in epidemiology requires efficient mathematical techniques. This paper compares two forecasting techniques for a time series data analysis: auto-regressive integrated moving average (ARIMA) and recurrent neural networks (RNN). Several industries are using classical time series forecasting techniques, like ARIMA, for instance market share pricing for oil industry [11]. Other industries are using machine learning forecasting techniques, for instance forecasting energy consumption [6]. So two most popular and widely used techniques from other industries have been chosen for a comparison.

The outline of this article is as follows: first, I describe the motivation, then I define the problem and set a goal for the paper. In Section 2, a brief review of literature on mathematical models describing disease outbreaks, the main techniques and the limitations of current approaches are identified. Basic notions of antibiotics resistance are presented. Section 3 presents two models for comparison and describes the details of the simulations setup. It discusses the role of machine learning techniques in inferring accurate predictors from observed data. Section 4 presents the results acquired from the model application to data of the drug-resistant bacterial infections. Finally, in Section 5, the findings are summarised and a view into future research direction is presented.

1.1 Motivation

Forecasting is the process of making predictions of the future based on past and present data. Forecasting of future behaviour in epidemiology is critical for aiding decision-making by public health officials, commercial and non-commercial institutions. The comparison of classical time series forecasting techniques and machine learning techniques has been done for the stock market analysis [12]. It was observed that the machine learning model Bidirectional Long Short-Term memory (BiLSTM) outperformed ARIMA. Additionally, a comparison of the models has been performed for an epidemiological data of influenza in Japan and US [13]. The combination of several machine learning algorithms showed consistent performance improvements. The outcome of all these studies has been somewhat mixed, but overall neural networks tended more to outperform classical linear techniques [1].

1.2 Problem Definition

The epidemiological forecasting problem is defined as a time series problem. The time series is acquired at discrete times with constant time lag Δt such that any acquiring time is $t_i = i\Delta t$, with $i \in \mathbb{N}$. In the following without loss of generality we set $\Delta t = 1$ and then $t = i$ such that $t \in \mathbb{N}$ and the time series results to be defined by $X_t = \{x(t = i) = x_i \mid i \in \mathbb{N}\}$. The forecast can be written as $\hat{X}_{N+h|N}$ which is the forecast h-steps ahead given the training set N . From above mentioned reasons, these forecast issues have spurred the need for forecasting in order to use the full capacity of the available data and be able to plan response activities.

1.3 Goal

The paper focuses on the analysis of medically important drug-resistant bacterial infections with data gathered by Antibiotics Resistance Surveillance from Robert Koch Institute, Germany [9]. The objective of this analysis is to compare the accuracy of the classical time series forecasting method with the machine learning forecast model, and use that information to determine when resistance will rise or decrease in order to develop an operational plan. The laboratory data were retrieved from January 2008 to October 2018, summed per quarter. The total number of data points is 44. The forecast has been done for 9-n steps ahead, what is equivalent to a two years forecast.

Here I am running an experiment verification of the proposed methods for forecasting drug-resistant bacterial infections. The results are conducted using five bacterial pathogens with increased resistance to commonly used antibiotics to compare accuracy of the proposed methods for the forecast of antibiotics resistance in the real environment.

2 State of the Art

In the current section a brief review of the literature on different forecast methods is done by identifying the main techniques and the limitations of current approaches. The section starts with a short introduction into the problem of antibiotic resistance.

2.1 Biological Overview

The current antibiotic crisis can be considered as an evolutionary problem [10]. Bacteria have a remarkable capacity to adapt and evolve even under extreme conditions including in the Arctic or in boiling water (near ”black smoker” hydrothermal vents). Bacteria can adapt specific genes or even lifestyles in response to the environmental conditions. Antibiotic resistance has developed to every antibiotic in clinical use, with the resistance genes responsible disseminated globally [5].

The analysis of the infection data offers the capability to potentially forecast the risk of drug-resistant bacterial infections outbreaks. Various attempts were made to investigate the development of the antibiotics resistance using different mathematical models on the research objects. For instance, as it is shown in Figure 1: within the patient, in the hospital or in the community. It is my aim to look at the spread of the drug-resistant bacterial infections in the community level.

2.2 Models Overview

Over the past two centuries the forecasting modelling approaches changed from descriptive and dynamic to network and social analysis, as highlighted in Figure 2. The descriptive approach focuses on the general shape and curve prediction based

Figure 1: Levels of antibiotic resistance forecast. Dynamic forecast of the drug-resistant bacterial infection in the human, in the closed community like in the hospital or in the open community.

on the real-life data. The dynamic approach in computational epidemiology mainly focuses on models where the whole population is divided into different groups (of susceptible, infective and recovered), and the transition among groups is modelled by differential equations [13]. Such models have limited prediction power due to lack of the ability to model individual level information [13]. The network approach is based on the understanding why the outbreak is happening and the implications on the networks and individuals. The social approach tries to forecast the risk of a disease for an individual.

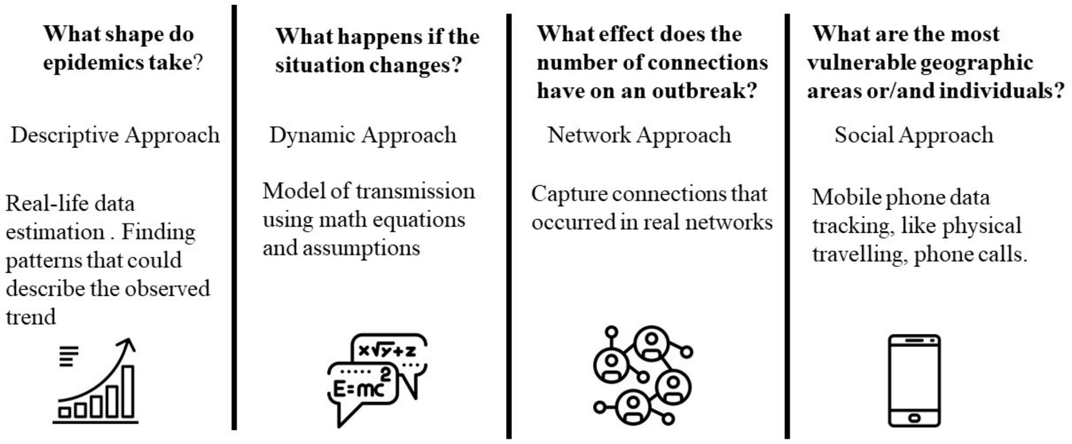
What shape do epidemics take?
Descriptive Approach
Real-life data estimation . Finding patterns that could describe the observed trend
What happens if the situation changes?
Dynamic Approach
Model of transmission using math equations and assumptions
What effect does the number of connections have on an outbreak?
Network Approach
Capture connections that occurred in real networks
What are the most vulnerable geographic areas or/and individuals?
Social Approach
Mobile phone data tracking, like physical travelling, phone calls.

Figure 2: Various approaches of infection disease modelling. The descriptive approach is based on the real-life data estimation. The dynamic approach is based on modelling the disease development via mathematical equations. The network approach focuses on the connections of individuals and groups. The social approach takes additional data like mobile phone data as input. Every approach has a goal to answer different questions in respect of the infectious disease development.

Here I am focusing on forecasting the spread for drug-resistant bacterial infections at the community level using the descriptive approach. My study investigates the patterns that could describe the observed trend based on the real-life data. So the epidemiological problem of the future number of cases of the drug-resistant bacterial infection can be defined as a time series problem.

Definition 1 *A time series is a sequence of observations x_t, where $t \in \mathbb{N}$ collected at equal spaced, discrete time intervals.*

Forecast of time series data is to model the future state as a combination of past data points [13]. The classical model considers the future state as a linear combination of past data, whereas machine learning models considers the future state as a non-linear combination of past data.

Classical Model. The general model for time series can be written as

$$x_t = g(t) + \epsilon_t;\ \ t = 1, ..., T \tag{1}$$

where T is the number of observations, $g(t)$ is a deterministic function of time, ϵ_t a residual term, or a noise, which follows a probability law.

In the time series analysis, it is assumed that the data (observations) consist of a systematic pattern and stochastic component; the former is deterministic in nature, whereas the latter accounts for the random error and usually makes the pattern difficult to be identified. The stochastic component of the time series is described by the error term ϵ_t. The basic assumption in time series analysis is that some aspects of the past pattern will continue to remain in the future [14]. It is necessary to evaluate if the time series components are time invariant, meaning the components are constant. It is assumed that the error distribution is the same by every data point: $\epsilon_t \sim N(0, \sigma^2)$, where N is the normal density with zero mean and variance σ^2. So, in this case the distribution still will be normal, and that its mean and the variance will still be the same 0 and σ^2. It demonstrates the invariant part of the process. So the goal is to find something invariant, which was the distribution of errors. The stationary series by definition have invariant parts such as unconditional mean and variance.

Machine Learning Model is an umbrella term for techniques that fit models algorithmically by adapting to patterns in data [8]. The prediction problem is defined as a problem of supervised learning problem. Possible non-linear dependence between the input (past embedding vector) and the output (future value) is attempted. The forecast is usually based on the idea according to which reliable predictions can be obtained solely on the grounds of our knowledge of the past. Chaos is often considered the main limiting factor to predictability in deterministic systems. [7, S. 567]

The representation of unknown input/output relation can be written as follows

$$\mathbf{y} = f(x) + \mathbf{w} \tag{2}$$

where $f(x)$ is a deterministic function, and the term $\mathbf{w}$ represents the random error.

The data are restricted to look like a supervised learning problem. The previous time stamp is an input variable x and the next step as the output variable y. Two hypotheses, which are seldom made explicitly, are needed to articulate an affirmative answer: 1. Similar premises lead to similar conclusions *(Analogy)*; 2. Systems which exhibit a certain behaviour will continue doing so *(Determinism)* [7]. The main limit to predictions based on analogues is not the sensitivity to initial conditions, typical of chaos. But, the main issue is actually to find good analogs [7].

Big data undoubtedly constitute a great opportunity for scientific and technological advance, with a potential for considerable socio-economic impact. To make the most of it, however, the ensuing developments at the interface of statistics, machine learning and artificial intelligence, must be coupled with adequate methodological foundations, not least because of the serious ethical, legal and more generally societal consequence of the possible misuses of this technology [7].

I have given a biological and models overview. For the next step, I am going to compare two models of forecasting the spread for drug-resistant bacterial infections at the community level: the classical model with ARIMA and the machine learning model with RNN.

3 Methodology

In this section the time series analysis of drug-resistant bacterial infection data will be used to make a forecast with two models: ARIMA and RNN.

The goal is to obtain a 9-step ahead forecast to test the forecasting performance of the ML method for a long horizon. Two methods are going to be compared on the same data set: the classical time series forecasting model ARIMA and the machine learning time series forecasting model RNN. The basic idea of ARIMA is to model the future state as a linear combination of past data points, whereas the basic idea of RNN is to model the future state as a non-linear combination of past data points.

3.1 ARIMA

In this section an overview of the theoretical explanation of the ARIMA model is given.

The current problem is defined as a regression problem, the output variable is a real value. The ARIMA model is a parametric method which represents data points as a linear combination of its previous values plus an error term. The approach is to model drug-resistant bacterial infection as time-series patterns with the ARIMA model. The model consists of three components: auto regressive (AR), integrated (I) and moving average (MA) models. As it was described above, it is important to evaluate if the time series components are time invariant, meaning the process is stationary.

An AR data model with p terms is constructed as follows, where p is the number of autoregressive terms included in the model; for a time series X_t

$$X_t = \mu + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + ... + \alpha_p X_{t-p} + \epsilon_t \tag{3}$$

where μ is the mean value, α_t is the weight of the correlation coefficients that is multiplied with the lagged values of X_t and t is the number of observations. The error term ϵ_t is an independent and identically distributed random variable from a normal distribution with constant mean and variance. The purpose of ϵ_t is to represent everything new in the series that is not considered by the past values. To fit an AR model to observed data, the order of the model, p, needs to be chosen and the parameters need to be estimated.

The I integrated data model has a d parameter, that represents the number of times that the raw observations are differenced, to achieve stationarity. A first order of differencing on the model would be written as

$$\Delta X_t = X_t - X_{t-1} \tag{4}$$

If the data set is stationary, then $d = 0$, and the model can be described as:

$$\Delta X_t = 0 \tag{5}$$

An MA data model is constructed as follows. The q is the size of the moving average window, also called the order of moving average.

$$X_t = \mu X_t + \epsilon_t + \beta_1 \epsilon_{t-1} + \beta_2 \epsilon_{t-2} + ... + \beta_q \epsilon_{t-q} \tag{6}$$

where $\beta_1, ...\beta_q$ are the sliding mean coefficients. MA is a modelling approach that uses the dependency between an observation and a residual error from a moving average model applied to lagged observations.

To combine the autoregression, integration, and moving average, the possible differentiation is made first, then the AR and MA equations are combined as ARIMA(p, d, q) and as ARMA(p, q), if d=0 :

$$X_t = \mu X_t + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + ... + \alpha_p X_{t-p} + \epsilon_t + \beta_1 \epsilon_{t-1} + \beta_2 \epsilon_{t-2} + ... + \beta_q \epsilon_{t-q} \tag{7}$$

In the Equation (7): p indicates autoregressive order; $\alpha_1, ..., \alpha_p$ are autoregressive coefficients; q indicates sliding average order; $\beta_1, ..., \beta_q$ are sliding mean coefficients; ϵ_t indicates a white noise sequence obeying normal distribution.

The basic modelling ideas and modelling steps can be summarized as follows: (1) making the non-stationary process a smooth process by means of difference; (2) establishing a suitable model to describe the stationary process; (3) predicting future values using the constructed model.

3.2 RNN

Recurrent Neural Networks (RNNs) are used as a deep learning framework to predict epidemiology profiles in the time-series perspective. RNN is adopted to capture the long-term correlation in the data [13]. The model is based on the standard sequence-to-sequence recurrent neural network architecture. RNN better fits to modelling problems such as time series data.

A neural network takes an independent variable **X** (or a set of independent variables) and a dependent variable **y**, then it learns the mapping between **X** and **y** (Training). Once training is done, a new independent variable can be given to predict the dependent variable.

The RNN has sequential input, sequential output, multiple timesteps, and multiple hidden layers. The Figure 3 highlights how RNN works. I calculate hidden layer values not only from input values but also previous time step values and Weights (W) at hidden layers are the same for time steps.

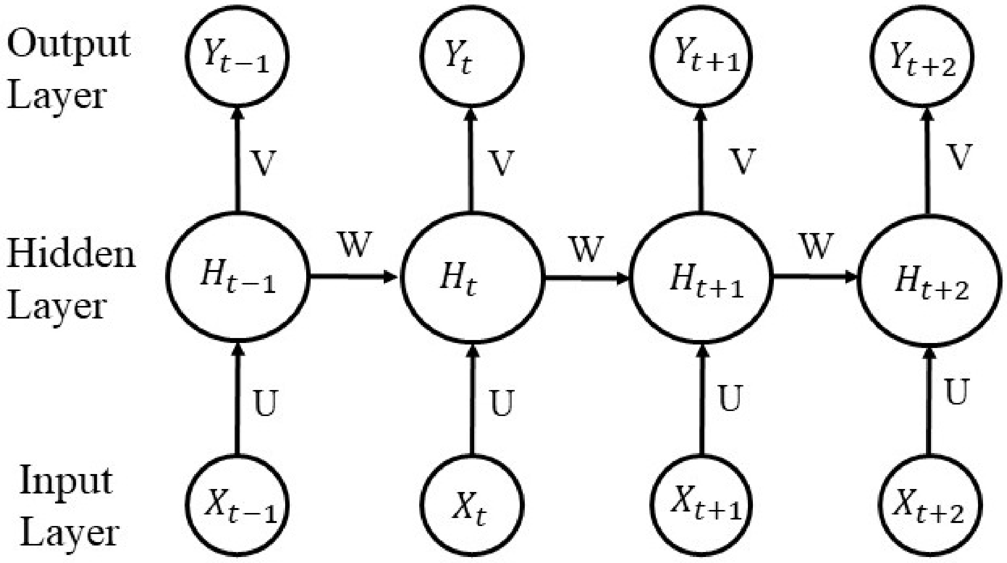

Figure 3: Description of the RNNs. U- Weight vector for hidden layer, V-weight vector for output layer, W - same vector for different time steps, X- Infection occurrences vector for input, Y- Drug-resistant bacterial Infections for output

The steps for the forecast are described described below.

3.3 Steps in Forecasting

The forecasting process involves the choice of the model, data splitting, fitting the model, model evaluation, re-fitting the model on the entire data set and forecast of the future behaviour. Figure 4 highlights the steps for a forecasting model. Each step will be further described. The output of the forecast is the number of drug-resistant bacterial infections towards a specific antibiotic at each time step.

Data Split. By convention, the test interval does not exceed 20% of the available data. So if the total available data set has 44 data points, then the test data set is 9 data points, which is equivalent to 9 quarters or 2 years. The training data set consists of 35 data points, what is approximately 9 years of observation.

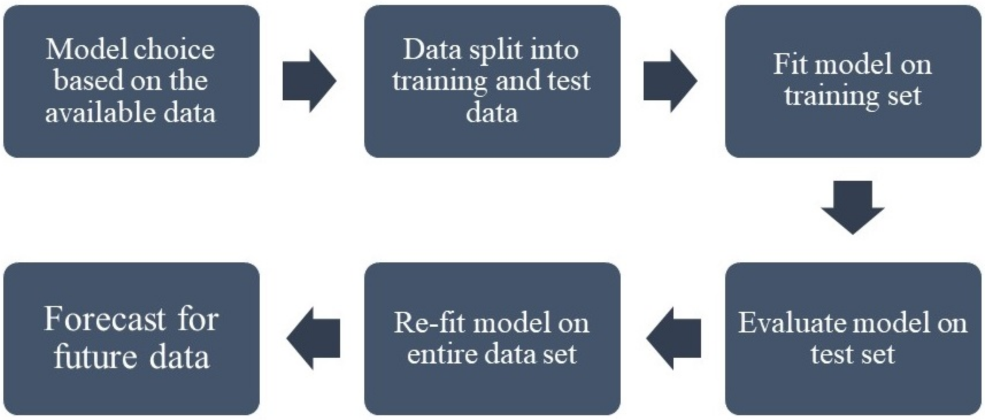
Model choice based on the available data
Data split into training and test data
Fit model on training set
Evaluate model on test set
Re-fit model on entire data set
Forecast for future data

Figure 4: The forecasting process consists of several steps: 1) the choice of the model based on the available data, 2) data split into training and test data sets, usually with the ratio 80:20, 3) the model is fit on the training data set, 4) model evaluation with RMS error, 5) the model is re-fitted on the whole data set, 6) forecast for future n-steps of data points.

If the dataset is denoted as $x_1, x_2, ..., x_N$, then the training dataset has the length T= 35 and denoted as $x_1, x_2, ..., x_T$. The test dataset has the length $h = 9$ data points and denoted as $x_{T+1}, ..., x_N$. In this example the test dataset and forecast dataset are equally long.

Fit model on training set. The parameters for the ARIMA model need to be chosen correctly, for (p, d, q). The parameters depend on the individual data characteristics. The RNN model requires weight function estimation to be performed. The model is initially fit on a training dataset or trained, meaning that the model is learning from the training data set.

Evaluate model on test set. The new segment Y_t is obtained with the ARIMA model. Y_t has the same length as the test data set X_t. During the process of prediction, the expectation is to obtain the best prediction value with no error or the smallest possible error, that is, to obtain the best future prediction value using root mean square error minimization. Root mean square error (RMSE) is defined as the root mean square of the ARIMA prediction and the actual value, during the time interval h, which denoted as

$$\lambda = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(X_t - Y_t)^2} \tag{8}$$

where Y_t indicates the predicted value, X_t indicates the true value from the test data set and N is the sequential number of steps in time in the time interval h.

Re-fit model on entire data set If the dataset is denoted as $x_1, x_2, ..., x_N$, the training data is set as an entire data set $x_1, x_2, ..., x_N$.

Forecast for future data The prediction interval should not be longer than the test interval. So the prediction interval is 8 data points, 8 quarters or 2 years. The model is trained on the entire data set of the length N. So the training data set should be at length at least N=44, for a forecast of h =8 data points - quarters - ahead.

Two time series forecasting models and steps of the forecasting process have been described. An experimentation comparison of the proposed methods for time series forecasting of the drug-resistant bacterial infectious disease is presented in the next section. The implementation of the proposed methods for the forecasting and results are shown in detail.

4 Use Case: ESKAPE bacteria

I have tested the forecasting models on the data of drug-resistant bacterial infections. The code used to evaluate the concept can be found on the github in [3]. I have used data of five highly relevant bacterial pathogens known as ESKAPE.

4.1 Available Data

ESKAPE is an acronym encompassing the names of six highly important bacterial pathogens commonly associated with antimicrobial resistance. Medically important bacteria include the following pathogens: *Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa* and *Enterobacter species*.

This study was conducted using data exported from the Antibiotics Resistance Surveillance from the Robert Koch Institute. Laboratories from all of Germany report these data on a regular basis [9]. The data are not differentiated by the region. Data used in this study included the time (quarter), the number of tested pathogens with confirmed resistance to a specific antibiotic. The laboratory data were collected from January 2008 to October 2018, summed per quarter. Thus, the data utilized for the analysis included quarter time series of isolates of *E. faecium* with Vancomicyn resistance, *S. aureus* with Penicillin resistance, *K.pneumoniae* with Amikacin resistance, *A. baumannii* with Imipenem resistance, *P. aeruginosa* with Imipenem resistance. The time series of *Enterobacter* were not available.

The forecast is the multiple-step-ahead forecast, for 9-n future steps. Forecast accuracy was assessed for each each time series. The data were treated as individual time series, analysed and evaluated separately.

4.2 Implementation

Implementation is showed in detail on bacteria *A. baumannii* with Imipenem resistance. Further ESKAPE bacteria pathogens have been analysed with the same strategy. Results are featuring analysis of the data sets from all bacteria pathogens.

Data Pre-processing Preliminary descriptive analysis was conducted at first. My aim is to identify relevant features, such as autocorrelation, seasonal patterns, trends, and any other notable fluctuations. Each time series was evaluated to determine whether it was stationary (i.e. whether basic statistical properties such as mean and variance of the series remained constant through time). Initial data analysis was conducted via computational tests of basic descriptive statistics.

The properties of this dataset are as follows:

- This is a real life dataset acquired from Antibiotics Resistance Surveillance from the Robert Koch Institute. The process is realistic. The quality of the monitoring data is high.
- The dataset consists of 44 data points, which were collected over the years from Laboratories in Germany. The data set contains observations from several laboratories.
- Each data point indicates the number of drug-resistant bacterial infections summed during three months.
- The dataset is not stationary.
- The dataset is univariate - one variable data input.

Data Split. The training data set is 35 quarter data points. The test data set is 9 data points. The ARIMA model and RNN use time series data as input. The data are reported in Figure 5 for the bacteria *A. baumannii*.

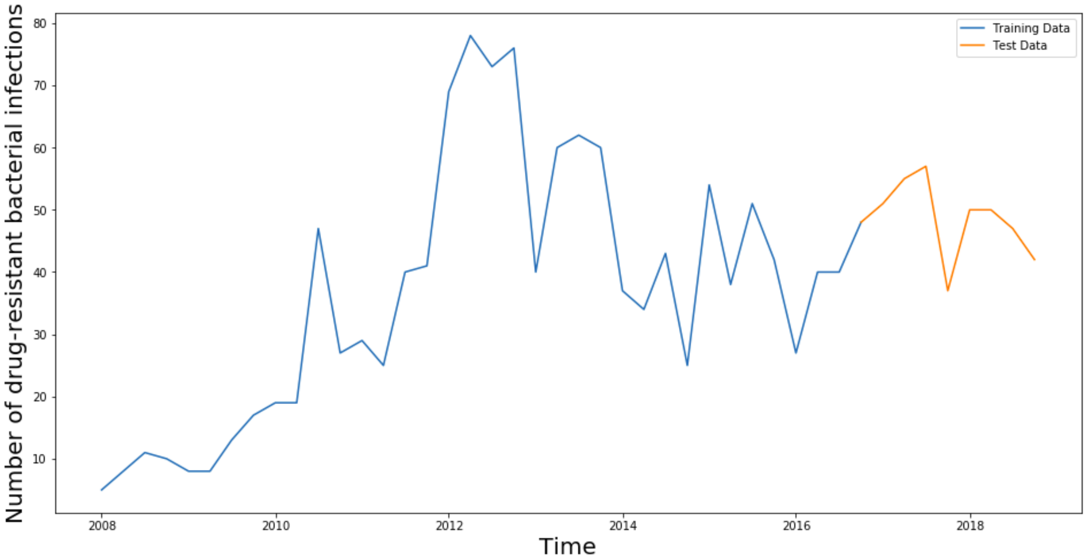

Figure 5: Drug-resistant bacterial infections from bacteria *A. baumannii* with Imip-inem resistance over the period of 10 years. The blue line represents the training data set of 34 data points and the orange line represents the test data set of 9 data points.

Fit model on training set. From the above analysis the data set is qualified to be modelled by ARIMA(p, d, q). The parameters have been estimated via computation $((p, d, q)) =$ (6,0,0)). These parameter values were also confirmed programmatically using the exhaustive grid search optimisation. Parameters have been estimated separately for each time series data set.

Evaluate model on test set. Forecast results, obtained with the ARIMA and RNN model are compared with the test data set.

The forecasts performed by both ARIMA and RNN models are reported in Figures 6-10 for the five ESKAPE bacteria. As it is seen from the figures, the RNN forecast does not replicate the changes as in the real data, whereas ARIMA does.

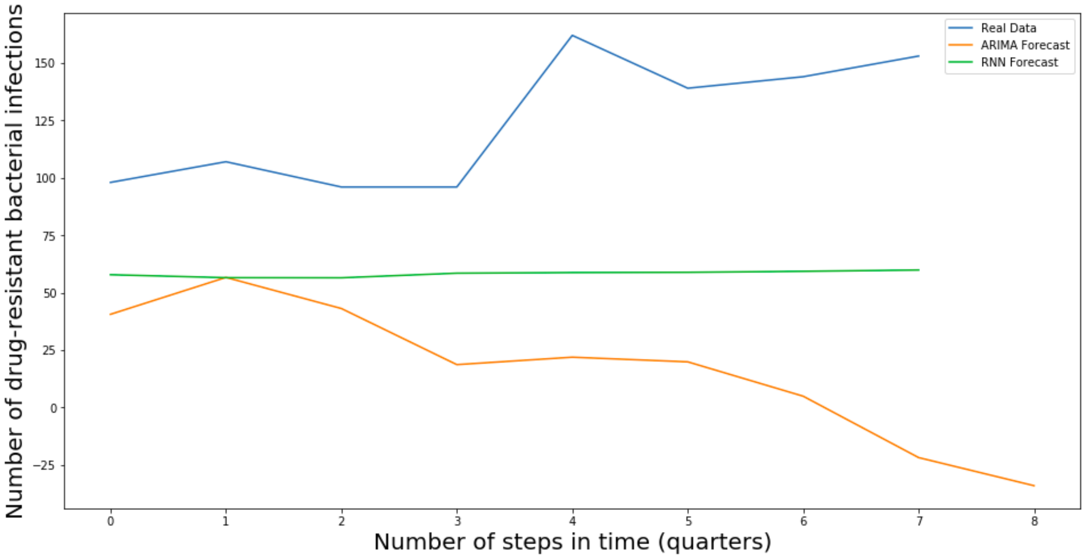

Figure 6: ARIMA forecast and RNN forecast of drug-resistant bacteria *E. faecium* with Vancomycin resistance. The blue line represents the test data set of 9 data points. The orange line shows the ARIMA forecast. Neither the ARIMA nor the RNN forecast have any common data points with the test data.

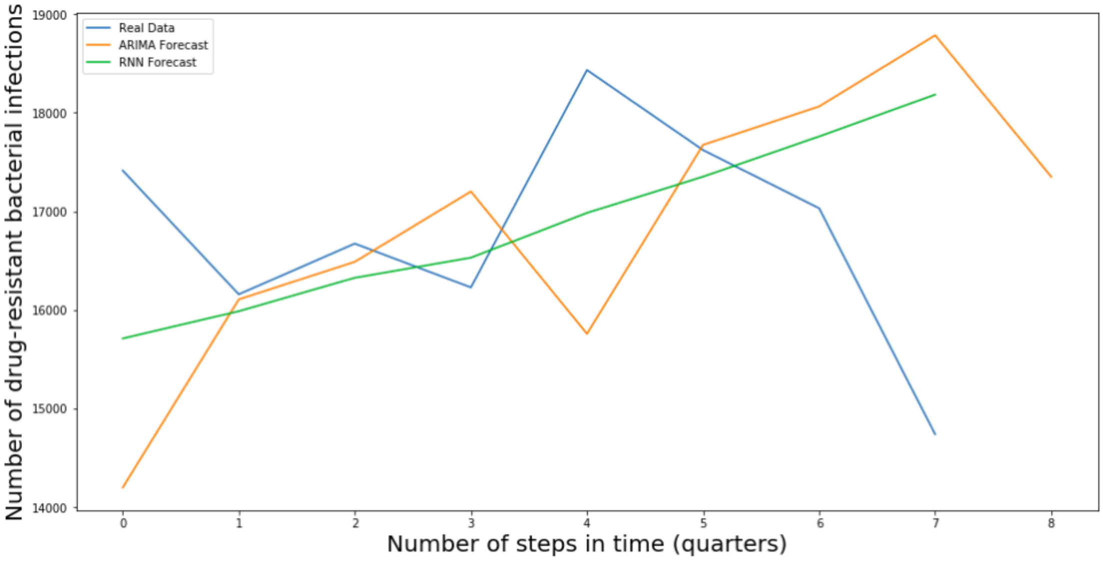

Figure 7: ARIMA forecast and RNN forecast of drug-resistant bacteria *S. aureus* with Penicillin resistance. The blue line represents the test data set of 9 data points. The orange line shows the ARIMA forecast. ARIMA has several common data points with the original blue line. The green line represents the RNN forecast, which also has several common data points with the test data.

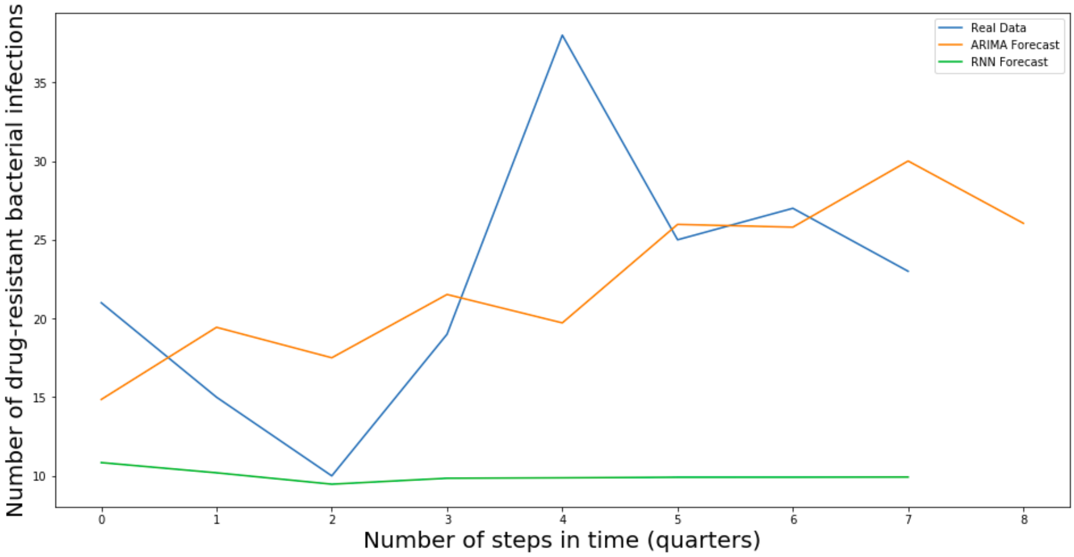

Figure 8: ARIMA forecast and RNN forecast of drug-resistant bacteria *K.pneumoniae* with Amikacin resistance. The blue line represents the test data set of 9 data points. The orange line shows the ARIMA forecast. ARIMA has several common data points with the original blue line. The green line represents the RNN forecast; it does not have any common data points with the test data.

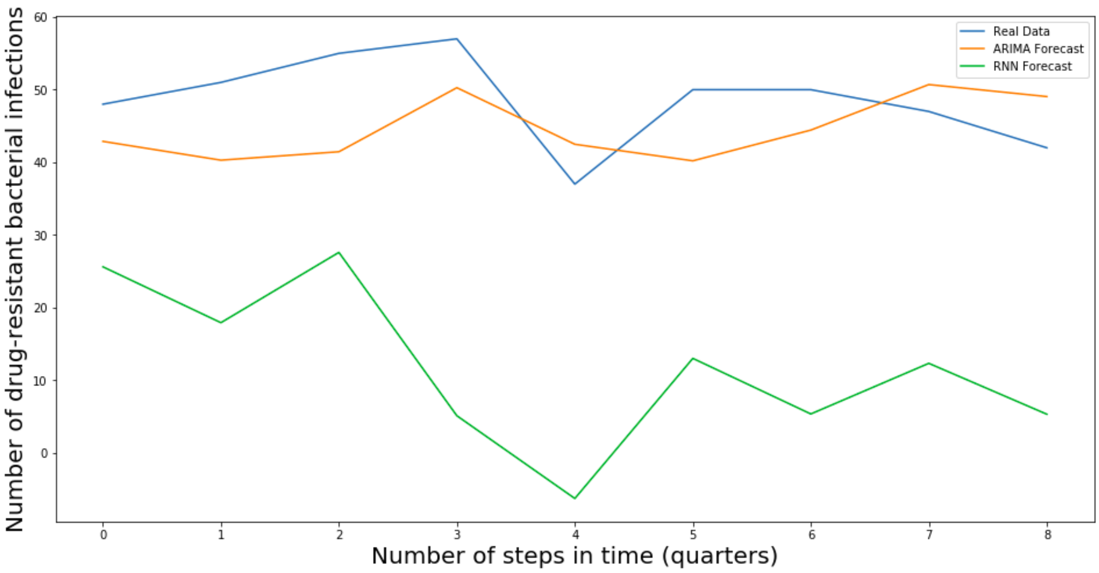

Figure 9: ARIMA forecast and RNN forecast of drug-resistant bacteria *A. baumannii* with Imipenem resistance. The blue line represents the test data set of 9 data points. The orange line shows the ARIMA forecast. ARIMA has several common data points with the original blue line. The green line represents the RNN forecast; it does not have any common data points with the test data.

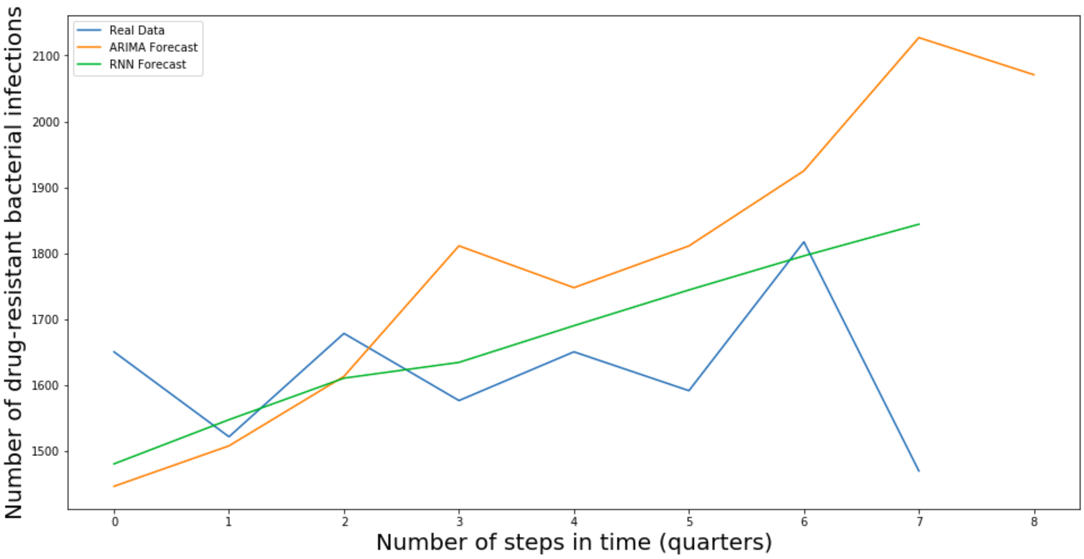

Figure 10: ARIMA forecast and RNN forecast of drug-resistant bacteria *P. aeruginosa* with Imipenem resistance. The blue line represents the test data set of 9 data points. The orange line shows the ARIMA forecast. The green line represents the RNN forecast; the RNN forecast has more data points in common with the test data than the ARIMA forecast.

4.3 Results

Two different time series forecasting method have been evaluated by the accuracy of the prediction.

Accuracy. Table 1 reports the Rooted Mean Squared Error (RMSE) achieved by each technique for forecasting the drug-resistance bacteria infection data. In three out of five test cases, ARIMA performs better then RNN. In particular, the RMSE for bacteria *A. baumannii* is λ =8.0 for ARIMA$(6, 0, 0)$ and λ =12.0 and for the RNN model. The similar difference in the RMSE is for the bacteria *K.pneumoniae.* The RMSE of ARIMA for the bacteria *S. aureus* is λ =1221.5, whereas the RMSE of RNN is λ =1416.6.

For two bacteria *E. faecium* and *P. aeruginosa* the RNN model has shown better results than the ARIMA model. The RMSE of the RNN model for the bacteria *P. aeruginosa* is λ =137.7, what is half of the RMSE of the ARIMA model with λ =273.9. The RNN model has performed significantly better for the bacteria *E. faecium* as well. As the data characteristics of the bacteria data are the same, it is hard to say why the results in the quality of the forecast differ so much.

Table 1: RMSEs of ARIMA and RNN

Bacterial pathogen	ARIMA	RNN
E. faecium with Vancomycin resistance	114.5	79.0
S. aureus with Penicillin resistance	1221.5	1416.6
K.pneumoniae with Amikacin resistance	5.9	12.5
A. baumannii with Imipenem resistance	8.0	12.0
P. aeruginosa with Imipenem resistance	273.9	137.7

4.4 Discussion

Our results suggest that machine learning methods do not always perform well as expected.

The observations describing the results are the follows:

1. Out of five time series data sets the ARIMA forecast model showed better performance for three data sets. The forecasting accuracy of the machine learning RNN model does not prove to be always significantly better as to that of ARIMA model for multi-step forecasting on univariate datasets. This work demonstrates that machine learning complexity is not always adding skill to the forecast. The potential reason is that RNN cannot preserve and thus does not remember long inputs [12]. Another hypothesis is that additional data preparation is required.

2. The RNN forecast model outperforms the classical ARIMA model for two data sets. As all five data sets have similar data characteristics, it is hard to say why the forecast accuracies are so different. This work demonstrates that there is no standard solution even for the similar datasets.

3. ARIMA requires the time series input only. This feature makes it relatively easy to implement, without high additional costs.

5 Conclusion

The purpose of this paper is to compare classical time series forecasting models with machine learning models on the real-life data of drug-resistant bacterial infection development. The classical time series models, such as ARIMA, models the future state as a linear combination of past data points. On the contrary, Machine Learning model such as RNN model the future state as a non-linear combination of past data points.

The paper was approached in the following steps:

Step 1. Define antibiotic resistance forecast problem at the community level as time series multi-step forecast problem.

Step 2. Check the available methods in the classical time series analysis and machine learning analysis.

Step 3. Forecast antibiotics resistance at the community level.

ARIMA and RNN were chosen as methods to forecast the number of drug-resistant bacterial infection occurrences with resistance. Forecast modelling was performed following the six steps: choose model, split data into train and test data, fit model on training set, evaluate model on test set, re-fit model on entire data set and forecast for future data.

Step 4. Experiment on the ESKAPE infection occurrences with resistance

The experiment included the data collection from Antibiotics Resistance Surveillance from the Robert Koch Institute for 10 years, as per the six steps indicated above. The results show that the ARIMA model performed well for three out of five data sets. Thus it can be seen that the RNN model does not perform well for multi-step forecasting on univariate datasets even when the data sets have similar properties. It can be explained by the fact that RNN cannot preserve and thus does not remember long inputs.

Future work can be concentrated on the combination of different machine learning methods and understanding in what cases the machine learning models deliver better results.

The results obtained by the forecast of drug-resistant bacterial infection occurrences are specific to a data set and should not be extrapolated to other data sets without previous verification. However the approach proposed in this paper for forecasting may be considered equally valid for other epidemiological data.

References

[1] Nesreen K. Ahmed et al. “An Empirical Comparison of Machine Learning Models for Time Series Forecasting”. In: *Econometric Reviews* 29.5-6 (2010), pp. 594–621. ISSN: 0747-4938. DOI: 10.1080/07474938.2010.481556.

[2] Anne E. Clatworthy, Emily Pierson, and Deborah T. Hung. “Targeting virulence: a new paradigm for antimicrobial therapy”. In: *Nature Chemical Biology* 3.9 (2007), pp. 541–548. ISSN: 1552-4450. DOI: 10.1038/nchembio.2007.24.

[3] Darja Strahlberg. *GitHub Repository.* URL: https://github.com/DarjaStrahl/AntibioticResistance_Comparison.git.

[4] Julian Davies and Dorothy Davies. “Origins and evolution of antibiotic resistance”. In: *Microbiology and Molecular Biology Reviews : MMBR* 74.3 (2010), pp. 417–433. DOI: 10.1128/MMBR.00016-10.

[5] Ayari Fuentes-Hernandez et al. “Using a sequential regimen to eliminate bacteria at sublethal antibiotic dosages”. In: *PLoS Biology* 13.4 (2015), e1002104. DOI: 10.1371/journal.pbio.1002104.

[6] G. E. Nasr, E. A. Badr, and and M. R. Younes. “Neural Networks in Forecasting Electrical Energy Consumption”. In: *FLAIRS-01 Proceedings, AAAI* (2001).

[7] Hykel Hosni and Angelo Vulpiani. “Forecasting in Light of Big Data”. In: *Philosophy & Technology* 31.4 (2018), pp. 557–569. ISSN: 2210-5433. DOI: 10.1007/s13347-017-0265-3.

[8] Stephen J. Mooney and Vikas Pejaver. “Big Data in Public Health: Terminology, Machine Learning, and Privacy”. In: *Annual Review of Public Health* 39 (2018), pp. 95–112. DOI: 10.1146/annurev-publhealth-040617-014208.

[9] Robert Koch Institute. *Antibiotika Resistenz Surveillance Datenbank.* URL: https://ars.rki.de/Content/Database/Introduction/Main.aspx.

[10] Roderich Roemhild and Hinrich Schulenburg. “Evolutionary ecology meets the antibiotic crisis: Can we control pathogen adaptation through sequential therapy?” In: *Evolution, Medicine, and Public Health* 2019.1 (2019), pp. 37–45. ISSN: 2050-6201. DOI: 10.1093/emph/eoz008.

[11] Ron Alquist, Lutz Kilian, and Robert J. Vigfusson. “Forecasting the Price of Oil”. In: *Bank of Canada Working Paper* (2011).

[12] Sima Siami-Namini, Neda Tavakoli, Akbar Siami Namin. “A comparative Analysis of Forecasting Financial Time Series Using ARIMA, LSTM, and BiLSTM”. In: *arXiv preprint arXiv:1911.09512* (2019).

[13] Yuexin Wu et al. *Deep Learning for Epidemiological Predictions.* 2014. URL: http://arxiv.org/pdf/1406.1078v3.

[14] Yufeng Yu et al. “Time Series Outlier Detection Based on Sliding Window Prediction”. In: *Mathematical Problems in Engineering* 2014.2 (2014), pp. 1–14. ISSN: 1024-123X. DOI: 10.1155/2014/879736.