# Predicting chickenpox cases in Budapest, Hungary

DSCI 725 – Sheyma Abdikebir

Introduction

In Hungary, the chicken pox disease has become an economic problem, imposing substantial financial strains on health care resources (Meszner et al., 2017). Hence, this project aims to apply predictive models on the chickenpox dataset to predict an outbreak or simply the future cases. These predictions allow hospitals and other healthcare institutions to devise effective strategies that can assist them in combating the impacts of the anticipated influx of cases, mitigating the significant clinical expenses caused by the disease in the country. Policymakers can use these predictions to inform their strategies. For instance, if an outbreak is anticipated, they can distribute their resources to health care institutions in a timely manner, helping them to better handle the impacts of the outbreak.

Background research

According to Pawaskar et al. (2021), chickenpox treatment requires patients to be quarantined for a week which can lead to lower levels of work productivity and even absenteeism in employees who contracted the disease or are caregivers, contributing to the economic burden caused by chickenpox. Additionally, Meszner et al. (2017) highlighted that although the expenses associated with curing a single chicken pox case is similar in Hungary, Spain and Germany, Hungary has a significantly larger healthcare costs per capita. Hence, the outcomes of this project will help healthcare institutions in Hungary to develop better resource allocation strategies and reverse the economic impacts of the disease. Policymakers can enhance the accessibility of vaccines by initiating campaigns in schools and care homes.

Problem presentation

This project aims to examine if the future chickenpox cases could be predicted in Budapest, Hungary. The objective is to use past chickenpox cases to make predictions with high forecasting accuracy.

Specification and design

Since the dataset is spatio-temporal, I utilized time series forecasting methods. Based on the results of my data exploration and the time-constraint, I focused on predicting the future cases in Budapest. Initially, the data were split into training and validation sets, then seasonal naïve was chosen as the baseline model to which I compared the performance of each model. Only those models that showed promising results were selected for model tuning. This report includes sections for descriptive statistics to present the results of my data exploration, predictive analysis to examine the predictive accuracy of each model, and graphics to visually illustrate and communicate the results of these models.

Data acquisition

This project utilized a dataset retrieved from Kaggle. It consists of the weekly chickenpox cases for 20 counties in Hungary starting from January 2005 and ending in December 2014.

**Chicken pox data**: <https://www.kaggle.com/datasets/die9origephit/chickenpox-cases-hungary>

Data exploration

The data set contained weekly chickenpox cases across 20 counties in Hungary. Initial data exploration suggested that the data does not contain duplicate and missing values. To define the scope of this project, I created a county-level bar chart to select the county with the highest chickenpox occurrence. The plot below showed that Budapest had the highest number of cases followed by Pest. Hence, Budapest was chosen as the primary focus of this project.

A graph of cases in each county

Description automatically generated

Moreover, the histograms below depicted that the cases in Budapest and the total cases across all counties were right skewed, indicating that high chickenpox cases were uncommon.

A graph of a distribution of cases

Description automatically generatedA graph of a distribution of cases

Description automatically generated

A graph showing the time series of chickenpox cases

Description automatically generated Furthermore, I created a timeseries plot in R for the total chickenpox cases in Hungary and for those in Budapest as shown below. Based on the plots, 2008 had a disproportionately higher number of cases. According to Szabó et al (2013), there were around a total of 900,000 chickenpox cases from 19 European countries in 2008, 2009, and 2010. The authors mentioned that this number highlights the widespread occurrence of the infection and the low rate of vaccine administration in Europe. This suggests that the recorded cases in 2008 were not an error owing to the high number of cases in those three years. Also, both plots showed that there is an annual additive seasonality because the variance does not increase with time. Moreover, the plot for the total cases showed a downward trend, but there is no discernible trend for the cases in Budapest.

A graph showing the number of cases in the same series

Description automatically generated with medium confidence

To explore the trend and seasonality, the Budapest time series data was decomposed as shown below. The additive decomposition showed that chickenpox is a seasonal disease. Budapest typically records its highest number of cases in winter and has fewer cases during fall and summer months. The trend component is unclear and shows fluctuations.

A graph of different types of time series

Description automatically generated with medium confidence

To better understand the trend in Budapest, a centered moving average with a window width of 52 was created because the data was recorded weekly, and it exhibited annual seasonality. Based on the plot below, it is evident that the weekly cases in Budapest displayed a constant trend with a slight decline starting from the end of 2010 through 2014.

A graph showing the original data and time

Description automatically generated

Data transformation

To ensure outliers were properly identified and replaced, Hyndman (2021) suggested the use of the function tsoutliers () on R that locates and replaces the outliers using nearby records through linear interpolation. However, since this dataset has seasonality, linear interpolation is not appropriate. In addition to the outlier mentioned earlier, the function identified another outlier in 2014 where the cases equaled 391. After examining the cases in other counties for possible outliers, the counties BACS and NOGRAD showed the same outlier in 2014 as illustrated below, suggesting that this value might not be an error. The outlier could be due to a possible outbreak in those neighboring counties.

A graph showing the time series of chickenpox cases

Description automatically generatedA graph showing the number of cases

Description automatically generated

The below plot shows the cases in BUDAPEST after replacing both the outliers with an average of the cases for their respective months. This transformed series was used to train and validate each model.

A graph showing the time series of chickenpox cases

Description automatically generated

# Theoretical Motivation for Choice of Technique

The initial set of models were chosen based on the types of forecasting models that are typically used by scholars for disease predictions. The training set consisted of 9 years’ worth of data starting from January 2005 to December 2013 and the validation set contained the 2014 cases. Each model was trained on the former set and its predictive accuracy was examined using the latter. Initially, mean average, naïve, and seasonal naïve models were created as candidate baseline models.

After analyzing the performance of all the models against the chosen baseline model, the two best performing models were selected for model tunning. The selection was based on the models’ predictive accuracy and the degree of overfitting with the goal of selecting those models with low errors. An iterative method was applied in which the parameters of the model were increased until it no longer reduced the errors.

**Note**: Since the chickenpox data has instances with zero cases which is common in disease related dataset, I did not incorporate MAPE as one of my error metrics to evaluate the models. This is because MAPE provides infinite or high error values when the actual series contains zero. Hence, alternative metrics were used, namely MAE, ME, RMSE, and MASE.

Iterative results

Firstly, I built 3 different baseline models, namely average method, naïve model, and seasonal naïve model. The seasonal naïve had the lowest errors compared to the other models. This is because the chickenpox data has seasonality. Hence, this model was chosen as the baseline.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **MASE** | **MAE** | **ME** | **RMSE** |
| Average method | 1.328433 | 57.84430 | -2.766506e+01 | 66.55774 |
| Naïve model | 1.1137242 | 48.49519 | 16.3413462 | 62.70260 |
| Seasonal naïve | 0.9079165 | 39.53365 | -24.543269 | 56.40250 |

A graph of a number of cases

Description automatically generated with medium confidence

A graph with numbers and lines

Description automatically generated

Next, Holt-Winter model, ARIMA, TBATS, and STL + ETS were applied on the training set. For disease prediction, Holt-Winter and ARIMA are usually adopted with the latter showing promising results. The plots are illustrated in Appendix 1. Visualizations of results. In this project, I also used STL + ETS and TBATS models which capture complex seasonality. These models were included because the residual plots of the ARIMA model showed autocorrelation at lag 104, indicating the presence of biyearly seasonality in addition to the previously determined annual seasonality.

Based on the table below, the Holt-Winter approach performed better than all the models, including the baseline model. The rest of the models did not perform significantly better than seasonal naïve. However, it should be noted that although ARIMA had slightly higher RMSE and MAE than seasonal naïve, it had the lowest MASE compared to all the models. This suggests that the ARIMA model can potentially provide promising results if tuned.

**Accuracy measures**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **MAE** | **ME** | **MASE** | **RMSE** |
| Seasonal Naive | 39.53365 | 24.543269 | 0.9079165 | 56.40250 |
| Holt-Winter | 33.66246 | 10.879473 | 0.7730806 | 41.71542 |
| ARIMA (auto.arima) | 42.94033 | -0.10226939 | 0.5402237 | 57.06257 |
| STL+ETS | 44.83282 | 29.3295795 | 0.9406565 | 56.08658 |
| TBATS | 45.72782 | 24.1252738 | 0.9594349 | 55.55416 |

Model tuning and Overfitting

The Holt-Winter approach and Arima model provided promising results, thus they were selected for model tuning. The former had relatively low RMSE, ME, and MAE values compared to other models, suggesting a lower degree of overfitting. ARIMA, on the other hand, had the lowest test and training MASE, suggesting a higher degree of prediction accuracy compared to naïve model. However, it should be noted that neither model captured all the information in the data as can be seen in the ACF plot of the residuals in the appendix

For Holt-winter’s method, I used manual grid search on R to find the best parameter values for alpha, gamma, and beta while minimizing RMSE to prevent overfitting. The parameters were allowed to vary from 0.1 to 0.9 with increments of 0.1. A coarse tuning was chosen as it requires less computational resources. The parameter values of the tuned model were as follows: alpha 0.2, beta 0.1, and gamma 0.4. While those for the initial model were as follows: alpha 0.2222126, beta 0.002289295, and gamma 0.6053701. As expected, the tuned model had lower MASE, MAE, ME, and RMSE than the initial model which can be seen in the appendix. However, the model did not capture all the patterns in the series.

For the ARIMA model, it should be noted that auto.arima does not always select the model with the lowest errors or BIC value. This necessitates comparing it with an ARIMA model whose parameters were manually selected. Hence, to tune the model, I used the ACF and PACF plots of the double differenced series to select the values for the non-seasonal term. From the ACF plot below, the cut off value is 1 and from the PACF it is 3. For the seasonal component, I started with no seasonal term then gradually increased both the P and Q terms. Since the series was double differenced both d and D were set to 0.

A comparison of a graph

Description automatically generated with medium confidence

The following table has been sorted from the best performing model to the least performing model. It is clear that the models with manually selected parameter values performed better than the model selected by auto.arima except for the model without the seasonal term which had the highest errors. The best ARIMA model was the model with the lowest errors, i.e. lowest degree of overfitting. After comparing all the ARIMA models with the model without the seasonal term, the greatest improvement was seen with ARIMA (3,0,1) x (0,0,1) model. This model had the highest predictive accuracy (MASE) and the lowest MAE, RMSE, and ME, suggesting a lower degree of overfitting.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **AIC** | **MASE** | **MAE** | **ME** | **RMSE** |
| ARIMA(3,0,1) x (0,0,1) | 4355.92 | 0.4864436 | 38.66555 | -0.07231792 | 52.98293 |
| ARIMA(3,0,1) x (1,0,1) | 4355.76 | 0.4893115 | 38.89351 | -0.1123293 | 54.86631 |
| ARIMA (3,0,1) x (0,0,2) | 4354.8 | 0.4974245 | 39.53838 | -0.1024751 | 56.19033 |
| Autoarima function =  ARIMA (2,0,1) (1,0,0) [52] with zero mean | 4398.3 | 0.5402237 | 42.94033 | -0.10226939 | 57.06257 |
| ARIMA (3,0,1) | 4456.57 | 0.7050034 | 56.03803 | 0.09755493 | 76.57160 |

Description of results

Compared to all the models, ARIMA (3,0,1) x (0,0,1) was the only model that successfully captured all the information in the data, resulting in a stationary ACF plot of the residuals. It should be noted that the tuned Holt-Winter model had the lowest RMSE and MAE out of all the models. For this dataset, ARIMA model proves to be successful for the following reasons:

* Although it has higher MAE and RMSE compared to the tuned Holt-Winter approach, it effectively captured all the patterns in the series.
* Produced lower MAE and RMSE compared to seasonal naïve and other models, except the tuned Holt-Winter’s method.
* Yielded the lowest MASE on training and test set.

These make the ARIMA model a satisfactory model. Hence, it is the winning model.

Applications and future directions

The forecasted chickenpox cases, as mentioned previously, can aid in formulating resource allocation strategies. To elaborate further, hospitals can use these predicted cases to provide adequate beds for patients, ensure the presence of effective diagnostic tools and medications, train volunteers, and hire enough medical practitioners. Additionally, these institutions can prepare patient educational programs in hospitals, including flyers and posters, ahead of time to prevent the spread of the disease.

To build upon this project, it is advisable to research the effectiveness of various strategies that can be adopted to slow the spread of the disease. These can include vaccination campaigns, quarantines, etc. Pairing the forecasts with the results of the research can help healthcare institutions to focus their efforts on strategies that yield maximum utility and thus optimize their operational planning strategies.

Limitations

Since this project focuses on predicting the cases in only Budapest, several limitations come to light. This project does not provide insights into future cases in Hungary (i.e. national level) Moreover, these forecasts are not generalizable to other counties. It should be noted, however, that if the cases in Budapest were correlated with the cases in other counties, then any forecasted outbreaks in Budapest can suggest that other correlated counties might have an outbreak as well. Finally, although these forecasts allow policymakers to allocate resources to healthcare institutions in a timely manner, the scope of the project does not allow them to do so at a large scale, i.e. including other counties, thus preventing efficient decision-making process.

Conclusion

To summarize, various models were assessed for their predictive accuracy. This was a necessary step in order to choose the best model for predicting the future chickenpox cases in Budapest, Hungary. Like the results of researchers in the field, ARIMA model performed the best as it was able to have low errors and high predictive accuracy (MASE). These predictions can be used to assist health care institutions, public health organizations, and policymakers devise data-driven decisions that can help them to minimize the impacts of the chickenpox cases in the economy. These forecasts can be used to alert the public of any potential influx of cases, motivating the community to take precautionary measures and reduce the spread of this highly contagious disease. Finally, this report discusses the future direction and limitations of the project.

APPENDIX

1. **Visualization of results**

A graph of different types of data

Description automatically generated with medium confidenceA graph showing the weather forecast

Description automatically generated with medium confidence

> accuracy(forecast, valid.tsp)

ME RMSE MAE MPE MAPE MASE ACF1

Training set -3.051223 45.87795 32.08860 -Inf Inf 0.7369360 -0.04292156

Test set 10.879473 41.71542 33.66246 37.39814 95.55547 0.7730806 0.02547699

Theil's U

Training set NA

Test set 0.9760428

A graph of a graph showing the number of residuals

Description automatically generated with medium confidenceA graph showing the time of a forecast

Description automatically generated with medium confidence

> accuracy(forecast, valid.tsp)

ME RMSE MAE MPE MAPE MASE ACF1

Training set -0.145229 49.24300 35.23738 Inf Inf 0.8092499 0.06637457

Test set -1.740982 39.06641 27.99498 -14.96888 68.44311 0.6429232 0.09689181

Theil's U

Training set NA

Test set 1.007021

A graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of

Description automatically generatedA graph showing a sound wave

Description automatically generated

> accuracy(pred, valid.tsa)

ME RMSE MAE MPE MAPE MASE ACF1

Training set -0.02262891 47.92747 34.00210 -Inf Inf 0.4277737 0.001167852

Test set -0.10226939 57.06257 42.94033 Inf Inf 0.5402237 -0.620671347

Theil's U

Training set NA

Test set 0

**A graph of a graph showing a number of residuals

Description automatically generated with medium confidenceA graph of a sound wave

Description automatically generated**

> accuracy(arima1.pred, valid.tsa)

ME RMSE MAE MPE MAPE MASE ACF1 Theil's U

Training set 0.11274341 51.82571 36.94855 -Inf Inf 0.4648424 -0.001404421 NA

Test set 0.09755493 76.57160 56.03803 Inf Inf 0.7050034 -0.620689615 0

**A graph of a graph of a graph

Description automatically generated with medium confidenceA graph showing the time of a sound wave

Description automatically generated with medium confidence**

> accuracy(arima2.pred, valid.tsa)

ME RMSE MAE MPE MAPE MASE ACF1

Training set 0.21735009 44.37218 31.97344 NaN Inf 0.4022515 0.0005006214

Test set -0.07231792 52.98293 38.66555 Inf Inf 0.4864436 -0.6251237016

Theil's U

Training set NA

Test set 0

**A graph of a graph of a graph

Description automatically generated with medium confidenceA graph showing the time of a sound wave

Description automatically generated with medium confidence**

> accuracy(arima3.pred, valid.tsa)

ME RMSE MAE MPE MAPE MASE ACF1

Training set 0.2135138 43.92441 31.53145 NaN Inf 0.3966909 0.0003352552

Test set -0.1024751 56.19033 39.53838 Inf Inf 0.4974245 -0.6363967630

Theil's U

Training set NA

Test set 0

**A graph of a graph of a graph

Description automatically generated with medium confidenceA graph showing the time of a sound wave

Description automatically generated with medium confidence**

> accuracy(arima4.pred, valid.tsa)

ME RMSE MAE MPE MAPE MASE ACF1

Training set 0.2185302 44.03666 31.59489 NaN Inf 0.3974890 0.0003247043

Test set -0.1123293 54.86631 38.89351 Inf Inf 0.4893115 -0.6359714853

Theil's U

Training set NA

Test set 0

A graph showing the time of a earthquake

Description automatically generated with medium confidence

A graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of

Description automatically generated

A graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of

Description automatically generated

> accuracy(train.tbats.pred,valid.msts)

ME RMSE MAE MPE MAPE MASE ACF1

Training set -0.2756978 42.92765 30.30243 -Inf Inf 0.6357881 -0.07867718

Test set 24.1252738 55.55416 45.72782 68.32948 138.4302 0.9594349 0.13771171

Theil's U

Training set NA

Test set 1.566667

> accuracy(train.stlm.pred,valid.msts)

ME RMSE MAE MPE MAPE MASE ACF1

Training set -0.4323936 38.34836 27.59441 -Inf Inf 0.5789701 -0.1045040

Test set 29.3295795 56.08658 44.83282 86.81035 148.7586 0.9406565 0.2115009

Theil's U

Training set NA

Test set 1.743949

**2: TASK TIMES: Predicted vs. Actual**

|  |  |  |
| --- | --- | --- |
| **Week 9 (March 3)** | **Predicted** | **Actual** |
| * Import the chickenpox dataset into R Studio and conduct initial exploratory analysis | 0.75 | 1 |
| * Perform data cleaning tasks such as locating and handling missing values, duplicates, outliers, and any other errors. Finally, implement data transformation | 2 | 2.5 |
| **Week 10 (March 11)** |  |  |
| * Implement various descriptive statistics on the data, create plots, and analyze the results | 2 | 3 |
| * Craft a list of potential forecasting models that will be applied on the data based on theoretical concepts | 1 | 1 |
| * Review midterm materials. |  |  |
| **Week 11 (March 18)** |  |  |
| * Split the dataset into a training and validation set. Perform initial analysis by applying the proposed methods individually and in combination on the training set | 3 | 4.5 |
| * Apply the model on the validation set, select the best performing models, and tune the parameters | 1.5 | 4 |
| **Week 12 (March 25)** |  |  |
| * Conduct final analysis and choose the best model out of the selected candidate models. | 1 | 2 |
| **Week 13 (April 1) N/A** | - | - |
| **Week 14** |  |  |
| * Prepare the first draft of the project data analysis | 2 | 3 |
| * Finalize the report, include diagrams and appendix section containing estimated and actual time taken to complete each task | 2 | 2.5 |
| * Submit the data analysis. | - | - |

**Note**: The above table is slightly different from the workplan. Initially, I planned to submit the proposal in Week 13. However, due to some unforeseen events, I started writing the draft in Week 14.

References

Hyndman, R. J. (2021, August 28). *Detecting time series outliers*. Rob J Hyndman. https://robjhyndman.com/hyndsight/tsoutliers/

Meszner, Z., Molnar, Z., Rampakakis, E., Yang, H. K., Kuter, B. J., & Wolfson, L. J. (2017). Economic burden of varicella in children 1–12 years of age in Hungary, 2011–2015. *BMC Infectious Diseases*, *17*(1). <https://doi.org/10.1186/s12879-017-2575-6>

Pawaskar, M., Méroc, E., Samant, S., Flem, E., Bencina, G., Riera-Montes, M., & Heininger, U. (2021). Economic burden of varicella in Europe in the absence of Universal Varicella vaccination. *BMC Public Health*, *21*(1). https://doi.org/10.1186/s12889-021-12343-x

Szabó, L., Jackowska, T., Kaló, Z., Kulcsár, A., Mészner, Z., Molnár, Z., Wysocki, J., Wutzler, P., Kormos-Tasi, J., & Sauboin, C. (2013, March). *Varicella vaccination in Hungary and Poland: Optimization of public benefits from prophylaxis technologies in the time of austerity\*\**. czytelniamedyczna.pl. <https://www.czytelniamedyczna.pl/4563,varicella-vaccination-in-hungary-and-poland-optimization-of-public-benefits-from.html>