Processing, Modeling, and Forecasting

A Time Series Analysis of Sick Leave Absences in Sweden and Evaluating the Impact of Macro Factors

Marcus Lindell, Casper Schwerin

Master's thesis
2023:E31

Lund University

Faculty of Engineering
Centre for Mathematical Sciences
Mathematical Statistics

Centrum Scientiarum Mathematicarum

Master's Theses in Mathematical Sciences 2023:E31
ISSN 1404-6342

LUTFMS-3475-2023

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lu.se/

Master Thesis

Processing, Modeling, and Forecasting: A Time Series Analysis of Sick Leave Absences in Sweden and Evaluating the Impact of Macro Factors

Casper Schwerin - ca4365sc-s@student.lu.se
Marcus Lindell - ma5323li-s@student.lu.se

MedHelp

LUND UNIVERSITY

Academic supervisor: Andreas Jakobsson
Industrial supervisor: Richard Sörberg
Centre for Mathematical Sciences
LTH, Faculty of Engineering
Sweden

Abstract

Sick absence affects companies both operationally and economically and the matter has become increasingly prominent following the COVID-19 pandemic. MedHelp Care is an e-health company working towards increasing workplace wellness by offering insights into absence and rehabilitation matters. Forecasting future absence levels could help with matters such as staffing and garner more insight into which factors that impact absenteeism. Moreover, gauging how external factors affect the absence could help deepen the knowledge further. With absence data now being digital, it is able to be modeled using statistical tools. The dataset was aggregated to a daily basis and preprocessed to better represent the Swedish labour force. Then, by applying time series analysis and machine learning methods, predictions of sick absence were formed. Datasets of the Swedish stock market and the discourse around COVID-19 on Twitter, respectively, were processed to time series and their relation to the sick absence was explored using crosscorrelation. The sets were then individually incorporated into exogenous models using autoregression, and then their impact was evaluated. Results show that the SARIMA model is better for predicting short and medium length horizons while the machine learning methods used were more apt for long horizons. Results from the exogenous models were mixed; the OMX set improved short term prediction accuracy while other horizons were largely unaffected, and the Twitter set worsened performance for all horizon lengths. While the model results are promising, their current complexity hinder large-scale deployment and further research is needed to further verify the effect of exogenous datasets.

Preface

This master thesis was conducted at MedHelp Care during the spring of 2023. We are very grateful for being allowed to become a part of the company and learn from different teams. In particular, we want to express sincere gratitude to Richard, our supervisor, who through his curiosity always encouraged us to push forward and helped us shape this project from the very start. Being a part of and contributing to MedHelp's goals concerning employee well-being was very rewarding and a great way to finish our respective master degrees within e-health. We found that playing table tennis is a great way of clearing our heads. Unfortunately, our colleagues did not find our playing quite as relaxing.

We also wish to thank Andreas Jakobsson at LTH, for helping us at every crisis, big or small. Without your input, we would still be on square one. Your book also proved to be quite the page turner.

Contents

Acronyms

MHB MedHelp Baseline
AR Autoregressive
MA Moving average
SARIMAX Seasonal Autoregressive Integrated Moving Average eXogenous input
WSS Wide-Sense Stationary
NP NeuralProphet
SGD Stochastic Gradient Descent
MAE Mean Absolute Error
MSE Mean Squared Error
MAPE Mean Absolute Percentage Error

1 Introduction

Work absence due to illness is a costly issue for work places and companies. The absence, generally characterized by frequency and duration, is the percentage of a work force on a given day that is made up by absent personnel. In recent years, an increasing amount of attention has fallen upon health in relation to the workplace. With matters such as the recent COVID-19 pandemic disrupting the status quo and forcing workplaces to reshape how and where employees may conduct their work, the subject of work absence has become more prominent. It has also become apparent that there are other factors apart from health that play a part in reported absence. Economical factors, such as the stock market, have been shown to affect hospitalization levels [1], while sickness benefits affects the size of the labour force [2]. What is more, the discourse in society, regarding sick absence has been explored by Swedish authorities in recent years and proven to influence absenteeism [3].

Many companies have started moving away from manual logging of absence and instead digitized the workflow. MedHelp offers an administrative tool for logging, managing and follow-up of work absence. They develop AI tools for detecting potential individuals showing signs of becoming long-term absent from work which has a negative economic impact on both individual and corporation.

There are numerous ways of understanding the work absence. One such method is by modeling, i.e., analyzing historic data which also enables prediction of future outcomes. In turn, approaching this data as a sequence of data over time, a *time series* [4], can be done in several ways. One such way is by transforming the series into stochastic processes that have properties that do not change over time. One then proceeds to detect potential periodic patterns in the data and adding parameters to the model. Another way of modeling a time series is through machine learning and instead approach the problem as fitting a function to historic data by minimizing errors. Using this procedure streamlines a lot of the workflow, at the cost of complexity. Whatever the approach, the systematic patterns included in models can then be used to form predictions of yet unknown data.

MedHelp currently has a tool for predicting calendar monthly averages of absence. However, they have yet to explore other granularities of the data concerning forecasting. To further garner insights into the matter with the purpose of increasing awareness to how absence will impact the current month, exploring time series forecasting with a daily resolution could also help facilitate staffing.

1.1 Purpose

The purpose of this project is to restructure existing absence data into a viable time series with a daily resolution and, using modeling methodology, garner further insights into the matter. This will be done by predicting future values, evaluating prediction accuracy and comparing results with existing and naive models. Additionally, exploring the data using a daily resolution, with the purpose of increasing awareness to how absence will impact the current month, could help facilitate staffing.

Furthermore, the intention is also to evaluate the impact of extending models to take exogenous input, factors which are believed to affect absence, into account and how it affects predictability. Given the current state of available packages in Python, the modeling possibilities are explored using both stochastic and machine learning models, which enables further comparison.

1.2 Limitations

While the long-term sick absence is generally more costly and may have more economic impact for companies, we have decided not to focus either on frequency or duration, as the data supplied lacks information about such matters. Additionally, this would require data on a more individual basis, which is somewhat contradictory to the statement of approaching the absence data on a population level. Although non-scheduled, data stemming from absence such as childcare is also dismissed, in order to render the work more focused on health related to the individual. Furthermore, we here limit our scope and exclude vacation, weekend, and compensatory leave, as these are not of particular interest when examining what drives work absence from a macro perspective.

This master thesis does not intend to find or explain any causality between work absence and macro factors, but rather how they might impact forecasting. In addition, we do not aim to predict future pandemics or their effect on absence. As a consequence, anomalies in the data, while not strictly speaking outliers in the stochastic sense, are addressed to reduce their influence and generalize the model in hope of better performance on "normal" data.

1.3 Report framework

The first part of the report describes the background to the project and the theory surrounding time series modeling. It is followed by a section describing the methods and processes that were used. Corresponding results are then illustrated before they are discussed in the last major section.

2 Background

2.1 Work absence

While the level of absence varies over the year, historically, the greatest non-scheduled reason for workplace absence is sickness [5]. It is defined by Hultin et al., as "absence from work due to illness", and also serves as a general method of gauging well-being [6]. The absence due to illness is often at its highest during the earlier months of the year and lowest during the summer [5].

While poor health and absence due to illness are strongly associated [7], the intrinsic link between the two is not an isolated entity, as there are numerous non-health related factors that tend to affect the workplace sick absence, often socioeconomic in nature [5]. Historically in Sweden, women generally have had a higher percentage of sick leave, $\approx$3.5 percentage points more in 2014. However, Statistics Sweden have identified this matter as declining over time [5]. Furthermore, studies have shown that gender plays a role for women regarding the risk of long-term absence, especially in workplaces where one gender is over-represented. Other factors could be events during the year, such as the olympic games, which cause elevated odds ($\times$1.46) of sickness absence [5].

In general, absence due to sickness have declined since 2000, with a 30% and 18% reduction amongst men and women, respectively [5]. Of the total sick absence (2000-2014), women account for 60% [5]. However, with matters such as COVID-19 occurring in the last couple of years, there is reason to believe a trend change might have occurred.

2.1.1 The MedHelp platform and dataset

Rather than manually noting which employees are absent on a given day, the process of reporting and recording employee absence has today been digitalized. Using the MedHelp platform, employees of companies can report their absence alongside additional information.

MedHelp Care is continuously developing and releasing AI tools for identifying individuals showing signs of becoming long-term absent, using factors such as gender, age, and previous absence history. Concurrently, they also offer companies insights into their respective workplace absence, e.g., predicted average calendar month absence as well as how they compare to other companies in the same industry. As a mean of broadening the knowledge regarding the larger population, currently over 200 000 users, the data can be approached in alternative ways. One such way is to explore data as a time series and analyze it with the potential of generating insights regarding trends and recurring patterns. Furthermore, the approach using the time series analysis with a changed granularity opens up the possibility of predicting future absence with a daily resolution. Such predictions could be used for staffing, i.e., ensuring the quantity of employees is sufficient enough for companies to operate adequately.

The dataset created for and to be used during this master thesis consists of approximately 6 million data points from Swedish companies and spans the period 2018-01-01 - 2023-02-14. Each point is labeled with calendar day, company ID, absence type, gender, as well as the number of individuals sharing the same

attributes. Each data point can therefore be seen as a subgroup of a company for each given calendar day and is illustrated in figure 1.

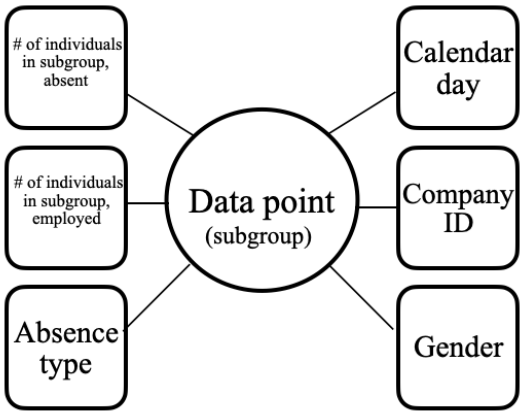

Figure 1: Illustration of a data point from the MedHelp dataset.

The structure of the dataset is a compromise between resolution and practicality. The lack of identifiers to individuals in the database restricts any analysis regarding frequency or duration, but the already grouped nature of the set renders it more convenient from an analytical perspective. As companies have joined the platform on different dates, the separation of data points amongst company IDs allows the data to instead be processed into absence quotients rather than counts of absent employees. Modeling the sum of sick absentees on a day would prove problematic, as the number would otherwise drift as more companies connects to MedHelp's platform.

MedHelp has an existing tool for forecasting absence which will be referred to as *MedHelp Baseline* (MHB). However, it is concerned only with forecasting the average absence percentage for a calendar month. At any given day of the month, it predicts what the average percentage will be for the current month. The model is a moving average, based on a combination of the of previous calendar month and the same month last year, as well as percentages observed during the month so far. A simplified version of how the algorithm weights data is shown in figure 2.

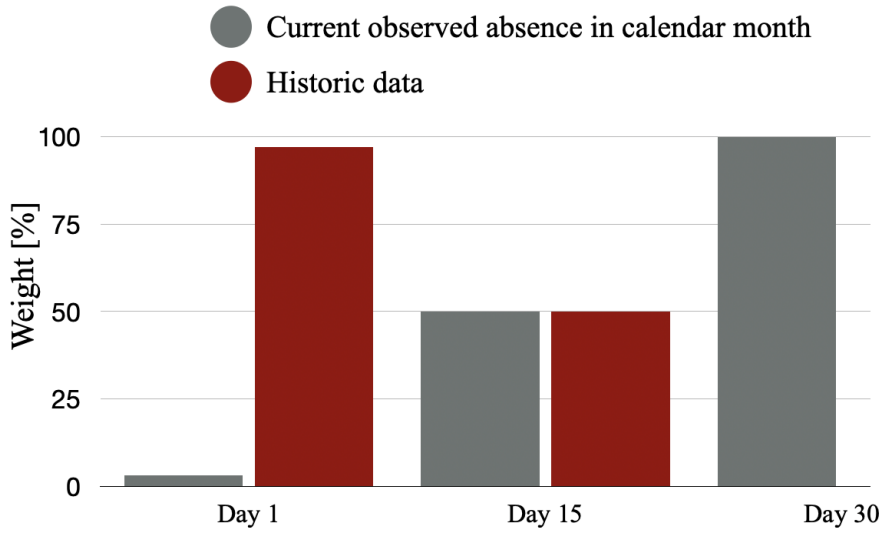

Figure 2: Simplified schematic of the MedHelp Baseline model. It emphasizes historic data early in calendar months and then shifts to recent data as the month progresses.

The current discourse around AI and its already strong presence at MedHelp Care informs the project that the previously mentioned time series analysis could be approached, at least partially, using machine learning methods.

Furthermore, there is an interest to investigate if there are larger external (*macro*) factors that affect modeling and that could potentially increase prediction accuracy.

2.2 Macro factors

While many of the socioeconomic factors that affect the sick absence concern workplace demographics and could be considered "local" in nature, matters at macro-scale are of more relevance when aiming to understand a larger population. For this report, social factors are separated from economic factors.

2.2.1 Online discourse

As some studies have investigated, traumatic incidents may affect absenteeism [8], i.e., the pattern in which employees are absent from the workplace [9]. On the other hand, there have been studies into how debate affects the levels of absence [10]. Whether or not the debate around workplace absence actually affects it has also been discussed by The Swedish Association of Local Authorities and Regions authorities [3]. The report also raises points about *social multipliers*, i.e., the effect one's close personal network may have on oneself and in turn how that might affect the network [3]. Investigating if the traumatic nature itself of the COVID-19 pandemic in combination with the social discourse affected the absence patterns is therefore of interest.

Measuring contemporary social discourse can be done using entities such as Twitter, which is one of the world's largest social media platforms with millions of users. The platform has been found to capture the real-time conversations and spread of information [11]. Furthermore, it has been found that archives from social media such as Twitter constitute reflections of societal discourse and impact [12]. The amount of data generated by the millions of users on Twitter each day, $>$ 100 million per day [11], renders it impractical to scrape, especially considering the rather large time frame. Instead, a preexisting dataset is used: TweetsCOV19 [12]. The set is minimally fragmented (4 segments) compared to other similar datasets and each tweet has a timestamp which is essential from a time series perspective. The set was originally limited to the time period March - April 2020 [12], but has since been updated to range from September 2019 - August 2022.

The dataset contains more than 41 million tweets where each post is connected to the pandemic through one of 268 keywords related to COVID-19, ranging from terms such as "coronavirus" and "lockdown" to "stayhomechallenge" and "SocialDistancing" [13]. The tweets have user ID:s and also a timestamp, making it possible to aggregate the number of tweets for a given calendar day.

2.2.2 Economy

A typical faucet of macroeconomics is to analyze and understand the market's relation to the household with respect to factors such as labour supply and household economy. We are here especially interested in the supply of labour as it is intrinsically linked to work absence. While wages play a significant role in the labour supply [2], it is not exhaustive. Income from nonlabor (e.g., return on savings) reduces both probability of entering the labor force and the number of hours which employees are willing to work. Furthermore, welfare schemes have a deterrent effect on labour [2].

The portion of GDP that can be accredited to absenteeism is estimated to exceed 2% within the European Union [14]. There has been a historical link between the larger economic situation and the levels of sick absence, as dire economical times seemingly caused higher discipline among workers [3]. While this relation was later dismissed following the global financial crisis of 2007-2008 [3], other economic matters surrounding sick absence has arisen. The same report also found that individuals that are denied sick pay are more likely to remain in work labour and work more hours [3]. In latter years, Swedish authorities have altered policies regarding paid sick leave; rather than no payment being made for the first day of sick absence (*karensdag*), 80% of the sick pay was paid (*karensavdrag*), starting January 2019 [15]. However, due to the COVID-19pandemic and to encourage employees to stay at home when ill, this reduction of sick pay was temporarily suspended during the periods 2020-03-11 - 2021-09-30 and 2021-12-08 - 2022-03-31 [15]. It follows that economy can be seen to affect the fluctuations of sick absence and could therefore be used as an analytical lens.

The size of the workforce in the MedHelp data set (average of $\approx$177 000 employees) informs that the economic factor used for exogenous modeling should be sufficiently large to be able to adequately represent the whole population. Stock markets can be used as a (causal) indicator of economic development,

both short-term [16] and long-term [17]. In 2021, around 2.3 million Swedish inhabitants owned stocks in Swedish companies [18]. One approach of representing the (macro)economy is therefore the Swedish stock market, more specifically OMX Stockholm 30 (OMXS30), which is the index for the top 30 companies when weighting according to the total market value of the shares **not** owned by their respective corporation.

2.3 Time series modeling

Before modeling, data should be checked for anomalies which are samples of data that differ substantially from others of the collection. Finding and correctly addressing these are key for striking a balance between model flexibility and the sporadic nature of real life data. There are numerous ways of detecting outliers which utilizes the variability of the data. One common method is calculating the standard score (or Z-score). This score is defined as the number of standard deviations (σ) past the mean value (μ) (above or below) that a given data point exceeds or surpasses. The z-value can be written mathematically as:

$$z = \frac{x - \mu}{\sigma}$$

where x is a given data point. To evaluate whether the sample is an outlier or not, a decision boundary is needed. A common value is $\pm$ 3, where values that are above or below are labeled as outliers.

2.3.1 Stochastic processes

To estimate the mean and correlation of a stochastic process, one has to impose some notable restrictions and assumptions on the process. Since the observation is treated as a realization, it has to exhibit some form of regularity. Thus assuming that the process is both stationary and ergodic, i.e., assures the statistics of the process will not vary over time and the characteristics of the process are measurable from only a single realization. These are rather strong assumptions and therefore one may need to examine relatively short segments of a time series to reasonably make the assumption that it is at least reasonably stationary. In particular, the processes of interest are so-called wide-sense stationary (WSS), implying that:

- The mean of the process is constant and finite.
- The autocovariance $C\{y_s, y_t^*\}$ only depends on the difference $(s - t)$ and not on the actual values of s and t.
- The variance of the process is finite, i.e., $E\left\{|y_t|^2\right\} < \infty$.

This makes it possible to further measure dependencies within and between WSS processes. When modeling time series, a popular approach is to describe the process as a moving average (MA), autoregressive (AR) or autoregressive moving average (ARMA) process. An MA process is defined as:

$$y_t = e_t + c_1 e_{t-1} + ... + c_q e_{t-q} \equiv C(z) e_t,$$

where $C(z)$ is a monic polynomial, i.e., its first coefficient is equal to one, of order q, $c_q \neq 0$ and e_t is a zero-mean white noise process with variance σ_e^2.

Subsequently an AR process is defined as:

$$A(z)y_t \equiv y_t + a_1 y_{t-1} + ... + a_p y_{t-p} = e_t,$$

where $A(z)$ is a monic polynomial of order p, $a_p \neq 0$ and e_t is a zero-mean white noise process with variance σ_e^2.

Combining the two, creates an ARMA process defined as:

$$A(z)y_t = C(z)e_t$$

The process is stationary if the roots of $A(z) = 0$ and invertible if the roots of $C(z) = 0$ lie within the unit circle, respectively [4].

2.3.2 Correlation functions

Identifying a model generally includes several steps in order to ensure that it appropriately describes the measurements and is thus considered a rather complex problem. The procedure is often iterative and follows the schematic in figure 3.

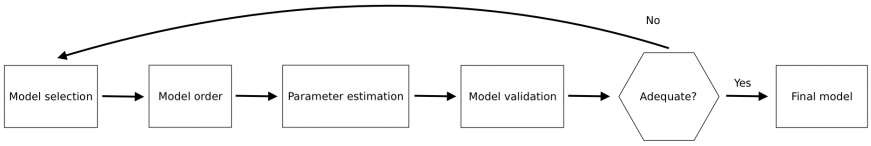

Figure 3: Schematic of the model selection methodology. The process is iterative and often require several attempts at selecting the model before results on the validation is deemed adequate and a final model is established.

If the model is sufficiently good, one may proceed, otherwise one has to iterate the procedure to come up with a model that is. Fortunately there are various tools and approaches for guidance when traversing the steps of identifying a model. Regarding the model structure, one may look at the autocorrelation function, ACF, and the partial autocorrelation function, PACF. Utilizing the ACF and PACF, allows for an easy way to determine the MA and AR lags, respectively.

Some systems may be described by additional measurements and thus it may also be of interest to look at the crosscorrelation between two WSS processes x_t and y_t.

2.3.3 Trends and seasons

The previously mentioned ACF and PACF can also give an indication of trends and seasons in the data, where a slowly decaying ACF implies a trend and a sig-

nificant peak in both the ACF and PACF at the corresponding lag characterize a seasonal trend [4].

Alternatively, one can perform an Augmented Dickey-Fuller (ADF) unit root test to evaluate stationarity. Considering an observed times-series, $\{y_t\}$, the ADF regression (with a time trend) can be formed as:

$$\Delta y_t = \alpha + \delta t + \beta y_{t-1} + \sum_{j=1}^{p} \rho j \Delta y_{t-j} + \epsilon_t, \quad t = 1, ..., T,$$

where $\Delta y_t = y_t - y_{t-1}$. Here ADF supports the inclusion of higher lags in contrast to DF. The hypotheses may then be formed as:

H_0 : Unit root, i.e., non-stationary,

H_1 : Stationary,

where rejecting the null hypothesis on a predetermined significance level implies that the series is stationary [19].

Before modeling, the trend and season have to be handled, either by removing it from the data and making it WSS or by integrating it in the model for it to be effective and precise. For this, the trend has to be identified, e.g., deterministic/constant, stochastic, or seasonal.

Handling a deterministic trend can be done using Fourier terms; treating it as a cyclic trend, similar to a seasonal trend. Fourier terms are a number of pairs of sine and cosine functions that are used to model seasonalities in the data. Fourier terms are useful for long seasonal periods, since they reduce the number of predictors that would have been used given a solution using dummy variables, which are more fit for short seasons [20]. Mathematically, a model using Fourier terms could be written as:

$$y_t = \alpha_0 + \sum_{l=1}^{d} \beta_l \cos(\omega_l t) + \gamma_l \sin(\omega_l t) + x_t$$

where ω_l denotes the lth frequency [4].

To handle a stochastic trend, one might instead integrate the trend in the ARMA model and extending it to an autoregressive integrated moving average (ARIMA) or even a seasonal ARIMA (SARIMA). An ARIMA process of order (p, d, q) is defined as:

$$A(z)(1 - z^{-1})^d y_t = C(z)e_t,$$

where d denotes the number of differentiations, in practice the data rarely requires more than two, $d \leq 2$.

Another way to integrate seasonal trends in comparison to the previous mentioned Fourier terms is to utilize a SARIMA model. A process may contain several seasonalities, e.g., weekly and/or yearly, thus a seasonal differentiation may also be reasonable. A multiplicative SARIMA process is defined as:

$$A(z)\mathrm{A}(z^s)\nabla^d\nabla_s^D y_t = C(z)\mathrm{C}(z^s)e_t$$

where ∇^d implies that the process is differentiated recursively d times, s is the seasonal period, e.g., weekly or yearly. The model orders for a SARIMA process is specified as SARIMA$(p,d,q)\times(P,D,Q)_s$.

While differentiation might prove sufficient to render the mean stationary, a transformation can be used to stabilize the variance of a process. One method for evaluating what transform might be required is the Box-Cox transformation, which regards maximizing log-likelihood for a parameter λ.

Commonly, it is sufficient to use a standard transformation given in table 1, by selecting the transformation closest to the maximized λ. This should be performed any other analysis such as differentiation [4].

Table 1: A selection of Box-Cox Standard Transformations. Given where the Box-Cox function is maximized, the table can be used to select the appropriate transformation for the data.

Values of λ	-2.0	-1.0	-0.5	0.0	0.5	1.0	2.0
Transformation	y_t^{-2}	y_t^{-1}	$y_t^{-1/2}$	$\log(y_t)$	$\sqrt{y_t}$	y_t	y_t^2

2.3.4 Predicting

The optimal linear k-step predictor for an ARMA process may be expressed as:

$$\hat{y}_{t+k|t}(\boldsymbol{\Theta}) = E\left\{y_{t+k}|\boldsymbol{\Theta}\right\}$$

where $\boldsymbol{\Theta} \equiv \begin{bmatrix}\theta^T & \mathbf{Y}_t^T\end{bmatrix}^T$, the parameter vector detailing the model of the process y_t and the measurements up to time t. Consequently the prediction error, $\epsilon_{t+k|t}(\boldsymbol{\Theta})$ may be formed as:

$$\epsilon_{t+k|t}(\boldsymbol{\Theta}) = y_{t+k} - \hat{y}_{t+k|t}(\boldsymbol{\Theta})$$

Considering an ARMA(p,q) process:

$$A(z)y_t = C(z)e_t$$

the process shifted k steps into the future may be expressed as:

$$y_{t+k} = \frac{C(z)}{A(z)}e_{t+k}$$

By expanding the polynomial division one may further express it as:

$$y_{t+k} = \left\{F(z) + z^{-k}\frac{G(z)}{A(z)}\right\}e_{t+k}$$

where $G(z)$ and $F(z)$ satisfies the Diophantine equation. Thus the optimal linear predictor can be written as:

$$\hat{y}_{t+k|t}(\mathbf{\Theta}) = \frac{G(z)}{C(z)} y_t$$

where $\hat{y}$ denotes the estimation of y [4].

2.3.5 Statsmodels

The statsmodels module in Python is an open-source software that provides classes and functions for the estimation of different statistical models, conducting statistical tests and statistical data exploration. The statsmodels TSA API provides tools and models designed for time series analysis such as ACF, PACF, CCF, ADF and SARIMAX for corresponding estimation. The SARIMAX class may be instantiated such that it enforces stationarity and invertibility when estimating the parameters of the model [21].

2.4 Machine learning

2.4.1 Neural networks

PyTorch is a library of functions for deep learning in Python [22]. It can be used to create neural networks consisting of layers of nodes, which are relatively structured to imitate the constellation of animal brains [23]. Neural networks are trained with input and their corresponding results and often optimized using algorithms such as stochastic gradient descent (SGD). SGD is an optimizer algorithm that estimates the local minima of a loss function using iterative computations of first-order derivatives (i.e., *gradients*) on a number of training data points (a *batch*) [24]. After each iteration, model weights, w, are updated, with a learning factor, γ, in accordance with the direction of the "steepest descent", in order to converge to the minima of the loss function, Q [25]. The equation can be formulated as:

$$w_{t+1} = w_t - \frac{\gamma}{n} \sum_{i=1}^{n} \nabla Q_i(w_t)$$

As proposed and designed by Triebe et al., neural networks can be used to mimic traditional regression for estimating AR coefficients and increase prediction performance, through a framework named *AR-net* [26]. Furthermore, by using regularization to penalize parameters, AR-net can automatically reduce the number of coefficients and consequently, approximating the correct AR order before coefficient estimation is not required [26].

2.4.2 NeuralProphet

In 2017, Facebook released Prophet, a tool for time series forecasting intended to be scalable for different applications [27]. Furthermore, the tool was made available in Python and R. Prophet transfers the time series analysis from dependence between temporal lags in data (i.e., AR and MA coefficients) and

instead poses the problem as fitting a curve to historic data. The creators argue that this approach increases interpretability and flexibility [27].

Four years later an independently developed tool named *NeuralProphet* was made available, indented as a successor to Prophet. This new model combined the Prophet approach to time series modeling with AR-Net [28], replacing the STAN backend with PyTorch. In addition to introducing neural networks to the task of time series forecasting, NeuralProphet also aimed to alter the close-ended nature of Prophet, allowing users to extend the library as they please.

Furthermore, NeuralProphet was created to address the problem of an ever-growing amount of data in combination with lacking overall resources and knowledge regarding time series [28].

NeuralProphet, like its predecessor, implements a model whose forecast can be split into additive components. Each module produces k (forecast horizon length) outputs, which are then element-wise added. The components are (all at time t):

- Trend: A combination of growth rate and offset. Using breakpoints, the rate can be changed an arbitrary number of times, effectively rendering this component a function made up of piecewise linear segments. Can be selected to be continuous or not.
- Seasonal effect: Seasons are modeled using Fourier Terms. The number of Fourier terms determines the robustness of the model: high orders may lead to overfitting and vice versa.
- Event and holidays effects: Binary dummy variables assigned to designated lags. With daily data points, this could be used for taking, e.g., sporting events into account when modeling.
- Autoregression effects: Allows the model, using AR-Net, to explicitly regress future values on earlier observed values with an order p, with a coefficient estimated for each order of included lag. The estimation of coefficients is done using a neural network with regularization, effectively rendering it sparse.
- Regression effects for lagged observations of exogenous variables: Time series which should in some way be related to the target series that is to be used for correlation. By serving as input signals, they could improve prediction accuracy.
- Regression effects for future-known exogenous variables: Other exogenous signals which have future values that are already known and can serve as prior information when forecasting.

The modular nature of the model originated from the desire to make the forecast more interpretable and provide local context [28]. Each component is modeled individually and then added to create the prediction, $\hat{y}$. Hence, in order to predict k steps into the future at a given time, each activated module needs to generate k values [28], by interpolating each individual components' dynamics.

When fed with data, a NeuralProphet model uses PyTorch to perform an optimizing function (e.g., AdamW, which extends SGD by including second order

derivatives), hence the data is divided into batches (which size, by default, is based on the size of the dataset). NeuralProphet supports a number of loss functions for estimating optimal parameter coefficients: Huber (default), mean squared error (MSE), and mean absolute error (MAE).

2.5 Evaluation and testing

The Akaike Information Criteria (AIC) is used for identifying and evaluating model selection. It estimates the prediction error using maximum likelihood and also penalizes the number of independent parameters k [29]. It can be written as:

$$AIC = 2k - 2\ln(\hat{L})$$

where a lower score signals a better model.

The Ljung-Box-Pierce test is a modified version of the Box-Pierce statistical test more apt for small sample sizes by including a scaling factor [30]. The test uses the ACF to test if there is structure left in model residuals (i.e., the quality of fit).

$$Q = N(N+2)\sum_{l=1}^{K}\frac{\hat{\rho}_{\hat{e}(l)}^2}{N-l}$$

Another test was proposed by Anna Clara Monti [31]. This test also evaluates if structure is present in model residuals, but by instead summing the squares of the PACF.

$$Q = N(N+2)\sum_{l=1}^{K}\frac{\hat{\phi}_{l,l}^2}{N-l}$$

The test score Q for each respective test may be compared to the χ^2-distribution with a selected significance level α and with K degrees of freedom. This can be formulated as

$$\chi_{1-\alpha}^2(K)$$

K is usually selected according to $15 \leq K \leq 25$ [4]. If $Q < \chi_{1-\alpha}^2(K)$, the null hypothesis that structure is present in the residual can be rejected.

Reviews of earlier modeling work regarding sick absence have shown that there is a scarcity of reports that evaluate model performance [32]. Some metrics that could be used to benchmark model prediction accuracy and statistical dispersion of residuals are

- Mean Squared Error (MSE) - in combination with variance, it offers insight into the potential bias of a model.

$$\frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$$

- Variance - A statistical measure of spread.

$$\frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$$

- Mean Absolute Error (MAE) - The average magnitude of the errors.

$$\frac{1}{N}\sum_{i=1}^{N}|(Y_i - \hat{Y}_i)|$$

- Mean Absolute Percentage Error (MAPE) - Scales error to facilitate comparison with other models.

$$\frac{1}{N}\sum_{i=1}^{N}\left|\frac{(Y_i - \hat{Y}_i)}{Y_i}\right|$$

2.6 Previous work

As is in line with the current goals of MedHelp's work, identifying individuals showing patterns that indicate future absence has been the focus of previous research. Research using the information gathered by companies in order to predict on an individual level has been conducted, where the frequency and duration of sick absence spells was the central focus [33]. Finding factors that aid when predicting was found to be necessary [33]. This was done using binary dummy variables to represent characteristics ranging from sociodemographic to job type as well as historic absence. While the study proved that historic absence was the strongest predictor, the other variables (administrative) still proved useful when creating models [33]. Historic short-term sick absence has been found to increase both future short and long-term sick absence [6].

Other studies, such as the one conducted by Zahid B Asghar et al., also used socioeconomic factors for constructing time series models [34]. The study found that statistical models could aid in forecasting future absence and in doing so facilitate with matters such as staffing and planning. Regarding COVID-19 specifically, one paper analyzed American data and modeled it while also incorporating Google Trends data and found promising results [35].

A large scale review of studies regarding modeling of sickness absence showed that most of the models are univariate, i.e., they only use historic data to predict future absence [32]. The review also found that the predictive accuracy of models were only evaluated in approximately 4 % of all selected studies.

In relation to economic factors, studies of historical data in California have shown a significant negative correlation between daily stock prices and hospital admissions [1]. The supposed relationship has a rapid effect, as plunging

stock prices incremented the number of hospitalizations the next 48 hours. The most prominent among these are mental health related, which in turn highlights a relationship between concerns about the future, more specifically consumption, and health. The study was conducted by estimating regression between hospitalizations and stock market return [1].

3 Method

3.1 Software

The vast majority of handling, processing, modeling and illustration of data is done in Python. To structure and define subtasks, Jupyter Notebook are utilized. By using a notebook interface, the readability of code is improved, as well as the efficiency, as certain segments of the code can be re-run rather than having to start from the top.

A major part of the necessary functions described in *Background* for SARIMA modeling are implemented from scratch or using combinations of functions from SciPy [36], scikit-learn [37] and statsmodels [21]. However, the lack of proper established intuitive tools for system identification, which would ideally facilitate the workflow in combination with the need of only storing data in RAM, renders modeling with exogenous input with statsmodels SARIMAX function rather unfeasible. Hence, exogenous input models are created using NeuralProphet.

3.2 Datasets of macro factors

OMX: Historic OMX data is downloaded from Yahoo Finance [38]. The metric selected is the closing value of each day when the stock market is open, as it is deemed the most respresentative. This dataset, due to its business nature, only contains data from weekdays. In order to render the time series compatible, the series needs to be interpolated. Values are created for indices that correspond to holidays and weekends using linear interpolation through *pandas.resample* and *pandas.interpolate*.

Twitter: As the TweetsCOV19 set simply contains tweets with information such as timestamps, each of the four sets are aggregated to number of tweets per calendar day. This is done by counting the instances of tweets on a given calendar day, based on their timestamp. They are then joined to constitute the whole timeframe, 2019-09-30 - 2022-08-31. As the TweetsCOV19 set contains tweets from each calendar day in range, interpolation is not required and no further processing is done at this stage.

3.3 Data preprocessing

Data is fetched from a remote repository and loaded directly into a pandas DataFrame. The data is divided into attributes such as company ID, gender, work type absence type and the corresponding number of individuals that belong to each such group. Any workers absent due to childcare is removed from the set, as they are of less interest since they are dependent on another individual. The data is then aggregated to correspond to a time series with daily resolution.

In order to ensure that the data is of high quality and more importantly representative of the Swedish labour force, the distribution of average absence per company is inspected in order to detect companies that need to be removed. There are then two types of companies, labeled as *atypical*, that should be extracted from the rest of the dataset:

- Companies with an abnormally high absence percentage

or

- Companies with an abnormally low absence percentage

By plotting each company that is present in the database against their respective mean absence during the period they were or have been present in the system, an overview of the data is given. While the companies with a high absence are removed using the outlier detection described earlier, those with exceptionally low percentages require more finesse. When illustrating the companies in the histogram, a substantial portion of the companies have an absence of exactly 0.0%, i.e., they have never reported any of their employees as absent. These companies will effectively dilute the total absence and are not seen as representative of the labour force, hence they are removed manually.

By dividing the absentees by the number of employees for a given day, the absence percentage per day is created, i.e.,

$$\frac{absentees}{employed}$$

The time series aggregation work flow is illustrated in figure 4.

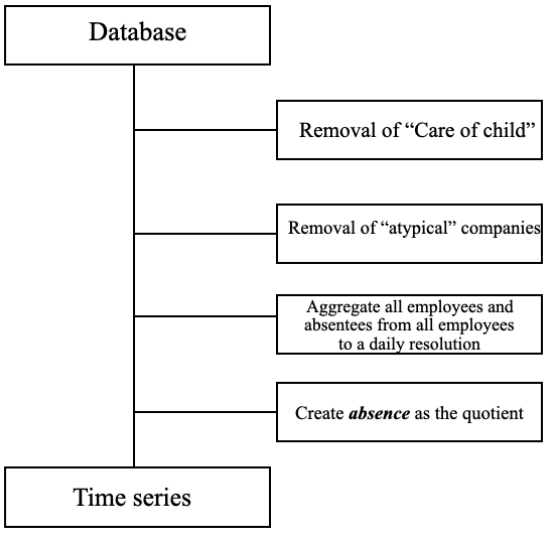

Figure 4: Preprocessing flowchart of the MedHelp dataset. The preprocessing reshapes the data into a time series that only constitutes sick absence from companies that have an average absence level within 3 standard deviations.

3.4 Modeling

In order to test model performance on unseen data, the last 15% of the set is chosen as a *test* set and consequently removed, while the remainder of the data is further processed. This series is analyzed for outliers, which are set to the appropriate threshold values ($\pm$ 3 standard deviations). The reasoning for this is that the time series have outliers that are very much in sequence and

setting these to, e.g., the mean value would hurt rather than help the model in the training phase. Once processed, a Box-Cox normality plot is used to check for a fitting transformation, in order to stabilize the variance. An ADF test is then performed to reject the null hypothesis, i.e., that a trend is present in the data. Finally, the set is split into *training*, 70% and *validation*, 15%, and their respective means are removed. The resulting data sets, as can be seen in figure 5, is used for all models. The only exception for this is the exogenous Twitter model. The same percentages are used for the sets, but as the TweetsCOV19 is shorter, it limits the length of the absence set. Consequently, the actual number of data points in subsets are different, which may affect the modeling procedure. This set is visualized in figure 6.

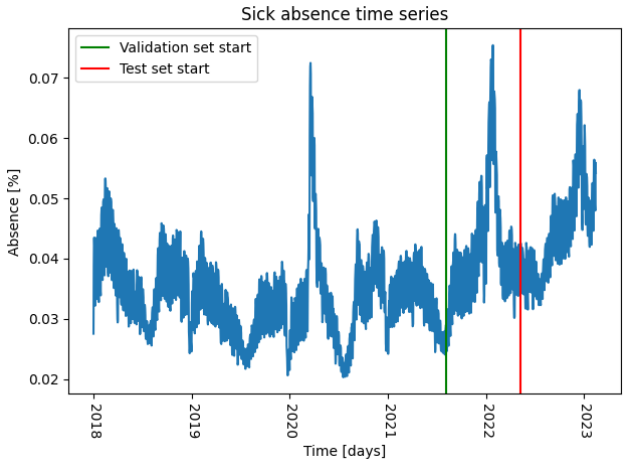

Figure 5: The time series is divided into three sets; modeling, validation, and test. The modeling set is used for estimating model parameters, the validation set for confirming the parameters, and the test set for evaluating performance on out-of-sample data.

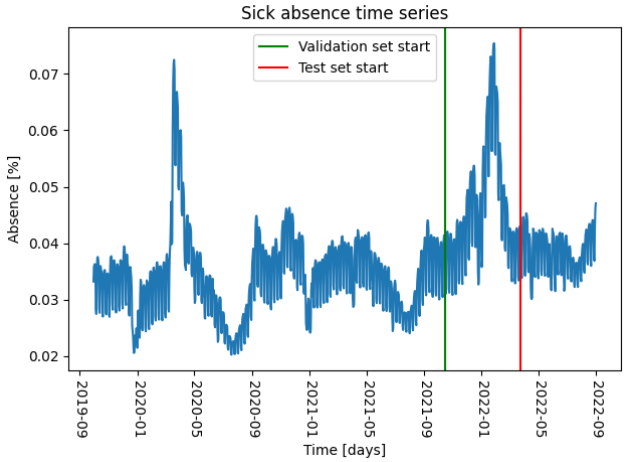

Figure 6: The time series intended for the exogenous Twitter model is divided into three sets; modeling, validation, and test. The modeling set is used for estimating model parameters, the validation set for confirming the parameters, and the test set for evaluating performance on out-of-sample data.

3.4.1 SARIMA

Once the data has been adequately preprocessed, the identification and modeling phase begins. Eventual trend and seasonality may be examined by looking at the ACF and PACF to provide an intuition of the data and what type of model structure that is required. When the model structure has been selected, the order of the model has to be determined. Analogously the ACF and PACF provides information about model order with the aim to come up with a model that renders a white noise residual. If the residual is considered white according to the Monti and Ljung-Box-Pierce test, the next step is to estimate the parameters of the model, i.e., the AR and MA polynomials. The parameters are estimated using the software which also provides the standard deviations of the estimations to conclude whether or not the parameters may be deemed significant.

When the models have been identified, with significant parameters that yields a white residual, a set of models may be evaluated with regard to the AIC score. However, the AIC score does not always reflect how well a model generalizes, i.e., the ability to adapt properly to new data. Thus, the models are also evaluated using the validation set where the best model will be used to form predictions of the test set.

3.4.2 Univariate NeuralProphet

NeuralProphet models requires only a rather small amount of preprocessing in order to be trained: indices, in this instance timestamps, needs to be renamed to ”ds” and the target value, sick absence, to ”y”.

Initial guesses of parameters are formed using random number generation (RNG), i.e., they are randomized based on an arbitrary number known as a *seed*. In

practice, this means that training results will vary slightly despite hyperparameters remaining unchanged. In order to ensure reproducability, the PyTorch seed is fixed, ensuring the same RNG and yielding the results to be consistent between instances. The validation data is used to evaluate performance of the model and detect potential over-/underfitting.

While NeuralProphet has default values for hyperparameters which enables usable models, these are intended for beginners to forecasting [28] and allows for plug-and-play. In pursuit of a more adequate model, which captures the characteristics of the data, the parameters need to be adjusted. A grid search enables testing of many combinations of discrete hyperparameter values by spanning a multidimensional room. In doing so, it allows for a streamlined process.

However, as the number of hyperparameters available in NeuralProphet are numerous, a selection is made before performing the grid search. Before committing to the very computationally demanding algorithm, the individual parameters are experimented with to evaluate their potential impact. Increasing the number of epochs swiftly led to overfitting and/or reduced impact per added epoch. Instead, a fixed amount of epochs, 50, was deemed a fitting number after manual testing.

The hyperparameters that are intrinsically associated with SGD are the batch size and the learning rate, hence they are included in the grid search. In addition, increasing the depth of the model with a number of hidden layers and number of respective units can assist in modeling complex dynamics are therefore included [39]. Furthermore, as COVID-19 is present in the training set, the weight of newer samples and the proportion of the data which is considered as "new" are tested, to evaluate how they might impact the training. Finally, three different loss functions are included in the grid: Huber, MSE, and MAE.

Each model that is a part of the grid search is trained with the same training set and the model residuals are then scored with a Monti test on both the training set and the validation set. A low score on the training set would infer that the model captures the dynamics of the model, while a low score on the validation set infers that the model performs well on unseen data. While the Monti test, especially for values of this magnitude ($Q \gg \chi^2_{1-\alpha}(K)$) is not a robust metric, it does provide insight to which of the models that capture the dynamics of the data. It is possible, however, for the models to ascertain a low Monti score on the validation set despite later displaying suboptimal results on unseen data. In order to avoid such issues for multivariate models, the MAE is also evaluated during the grid search to ensure that predictions and ground truth indeed are of similar proportions.

The number of combinations of hyperparameters results in a grid search constituting 648 unique models. Despite the fact that model residuals might not be deemed white, they can nonetheless overfit to behaviour such as the pandemic outbreak. Ideally, the model would perform similarly on the two sets, however, the validation set is more telling regarding performance on future datapoints.

The next step is to train the model with a set that is the combination of training and validation, with the same hyperparameters. Once the training is complete, the model is ready to create predictions for the different horizon lengths.

3.4.3 NeuralProphet with exogenous input

As previously stated, the lack of proper tools available for exogenous input modeling using SARIMAX models informs the project to only explore exogenous models using NeuralProphet.

The preprocessing of the exogenous sets for usage with NeuralProphet is handled in a different, simpler manner than the MedHelp dataset. As preprocessing the sets in a similar way to the sick absence diminishes their impact during modeling, they are instead scaled to a similar size to the absence set. The exogenous input in NeuralProphet is incorporated by extending the model with an additional additive component to account for AR coefficients of the input. A schematic of how exogenous input is integrated can be seen in figure 7.

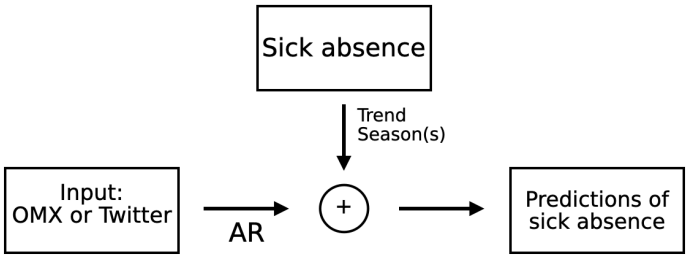

Figure 7: A schematic of how input and output are related when creating exogenous NP models. Trends and seasonality of the the target series (sick absence) are combined with AR coefficients from an input set such as OMX or Twitter. The components are then added to form predictions.

Searching for adequate multivariate models is othervise very analogue to the univariate method: NeuralProphet is trained with data and a grid search is conducted to find optimal model hyperparameters using Monti scores on modeling and validation data.

The grid search, in principle, is rather similar to the one conducted for the univariate case expect for a few changes. Some hyperparameters, such as the learning rate and loss function has had its search space tweaked and/or reduced as a result of preliminary testing of their effect on multivariate performance. Some parameters are new: *n lags*, previous lags of data to take into account (effectively AR-order) and growth type; continuous or discontinuous. For the exogenous models using OMX and Twitter, grid searches of 864 and 1024 models, respectively, are conducted.

The supposed optimal hyperparameters are then tested manually to see how they perform and what type of behaviour they exhibit (i.e., see that they adhere to the seasons of a normal year rather than fitting to anomalies). Once these matters have been inspected, the model can be set. For NeuralProphet prediction with exogenous input, the horizon length needs to be set before training.

3.5 Prediction

3.5.1 Predictions of monthly averages

While predictions made by the constructed models will be k-step (as in each step until k), it is possible to compare the models to the MedHelp Baseline model by averaging the k predictions as can be seen in figure 8. Furthermore, one can iterate and reduce the number of predictions as the number of observed percentages (data points) for the current month increases.

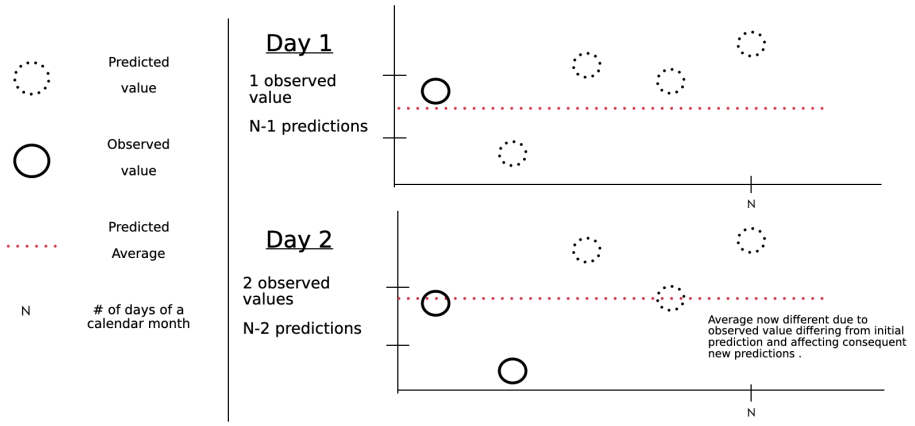

Figure 8: Iterative predictions of sick absence: Respective prediction means are calculated in each iteration, which is then used as the prediction on a given day of the average calendar month absence.

The nature of how data is fed to the MHB model means the model predicts with the same amount of data each month, i.e., its internal memory does not grow during the calendar year. The new data that is being added for each iteration of the prediction horizon could in theory be used to update and tweak model parameters. The re-estimation of these parameters are computationally demanding, especially given the model complexity. Furthermore, it would technically no longer be the same model between predictions which poses a dilemma regarding robust comparison with the naive and baseline models. In order to address this predicament, the model parameters for the SARIMA model are estimated using only the training set and then fixed. The new data that is fed then simply serves a method for "moving" the model forward. The hyperparameters for the NeuralProphet are given the same treatment; fixed according to the training set. Notwithstanding, the data which the NeuralProphet model is trained with grows over time which should, in theory, render it more accurate for each iterations of predictions.

Given the nature of the MHB model, the forecast horizons are strictly defined to calendar months. Consequently, the test set needs to be approached with this in mind. The test set can therefore be compartmentalized into 12 distinct sets, one for each calendar month of 2022. As the MHB model is slightly tweaked between calendar years, it would prove unfair to evaluate the model with a set that includes months from two different years. Hence, only the calendar months

of 2022 within the test set are used for the comparisons. Furthermore, the MHB model is dependent on the previous calendar month, and as the first full calendar month in the test set for univariate and Exo OMX is June 2022, the MHB comparison set for is defined as July - December 2022. By applying the same methodology to the TweetCov19 set, the corresponding months are May - August 2022.

3.5.2 Predictions of daily absence

Three different horizon lengths are considered and tested: 7 days (week), 30 days (month) and 90 days (quarter). In order to ensure a robust evaluation, the horizon lengths are iterated upon throughout the test set, see figure 9, which significantly increases the sample size and satisfies the approximation of normal distribution. Between each respective iteration, the NeuralProphet models are retrained, while the SARIMA model remains static. Only predictions from complete horizons are used which results in 40 week-long predictions (280), 9 month-long predictions (270) and 3 quarter-long predictions (270) for the univariate models and the exogenous OMX model. For the exogenous Twitter model, this instead infers 22-week long predictions (154), 5 month-long predictions (150), and 1 quarter-long prediction (90).

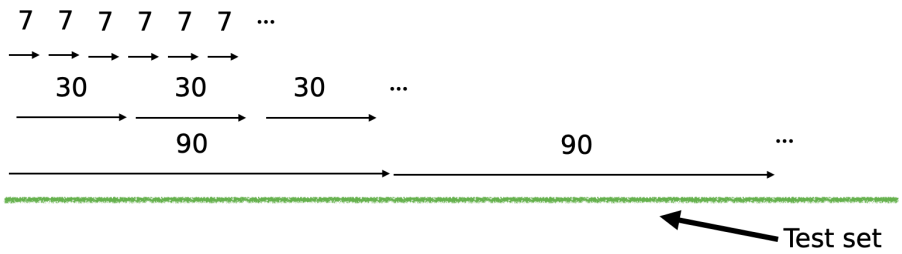

Figure 9: Iterative prediction horizons throughout the test set: The predictions within an horizon are formed and then the models predict the subsequent horizon until the end of the test set is reached. Note that only complete horizons are used.

In order to benchmark, the models are compared to a naive model, which forms its predictions as last year's sick absence, i.e., $\hat{y}_t = y_{t-365}$.

3.5.3 Prediction evaluation

When evaluating the predictions, it is feasible to do so in the original domain. Thus, all predictions have their respective mean re-added and are subsequently inverse transformed before evaluation. To measure performance of model predictions, four evaluation metrics are used: MSE, variance, MAE, and MAPE.

4 Results

4.1 Preprocessing

The preprocessing workflow and its numeric counterparts can be seen in figure 10.

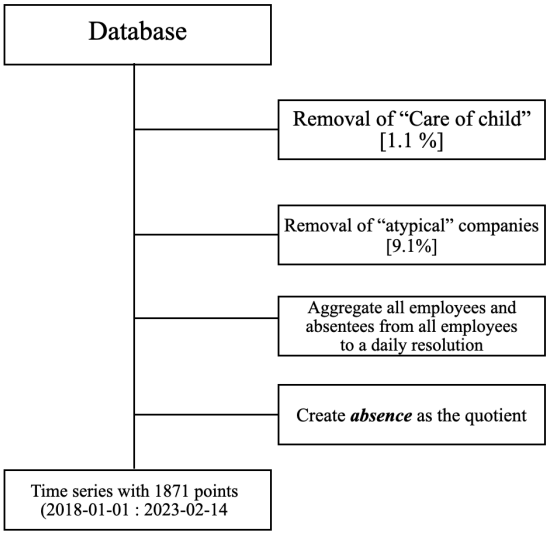

Figure 10: Resulting workflow of the preprocessing of the sick absence, where absence due to childcare and companies with unusual average absence levels are removed. The percentage of total data that was removed are shown in the respective steps.

The distribution of companies and their respective average sick absence after removing deviating companies can be see in figure 11. In total, 28 datapoints were detected as outliers.

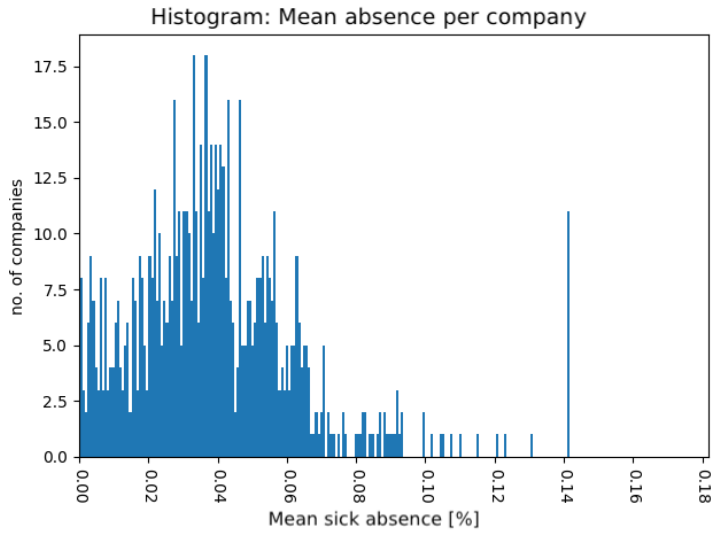

Figure 11: Histogram of companies' average sick absence. Here, outlier companies have been removed to better represent the Swedish labour force.

The corresponding difference in sick absence levels after removing these companies are shown in figure 12. The difference is most pronounced earlier in the series.

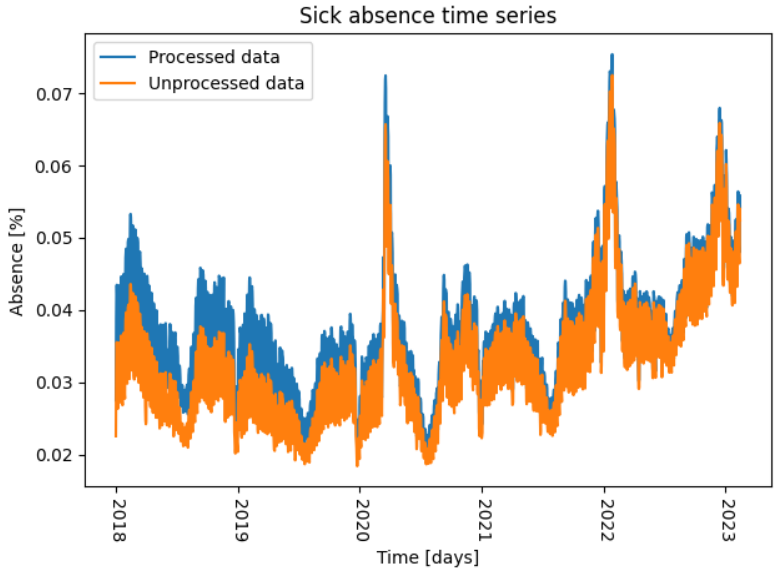

Figure 12: Daily sick absence from 2018-01-01 to 2023-02-14. The orange plot illustrates the absence data from all companies in the database (unprocessed) and the blue plot is processed data, where companies with abnormally high or low average absence levels have been removed.

For the dataset used for the univariate and exogenous OMX models, a log-transform is deemed appropriate by the Box-Cox transformation. For the exogenous Twitter model an inverted square root transform is instead suggested and applied to the absence set.

4.2 Model parameters

For the SARIMA model, a SARIMA$(2,0,7) \times (0,1,0)_7$ or:

$$A(z)\nabla_7 y_t = C(z)e_t$$

seems suitable with:

$$\begin{aligned} A(z) &= 1 - 0.2076(\pm 0.028)z^{-2} \\ C(z) &= 1 + 1.183(\pm 0.015)z^{-1} + 1.0514(\pm 0.023)z^{-2} + 0.9964(\pm 0.028)z^{-3} + \\ &\quad 0.9172(\pm 0.027)z^{-4} + 0.8325(\pm 0.022)z^{-5} + 0.8936(\pm 0.021)z^{-6} + \\ &\quad 0.0663(\pm 0.019)z^{-7} \end{aligned}$$

NeuralProphet model parameters are presented in table 2. The most significant differences between the univariate and exogenous models are the weight of newer samples and the AR order to apply on input datasets.

Table 2: NeuralProphet model hyperparameters for the different models. The hyperparameters were found using grid search. * denotes the default value.

Hyperparameter	**Univariate**	**X:OMX**	**X:Twitter**
Learning rate	0.05	0.01	0.01
Batch size	120	120	120
Number of epochs	50	50	50
Number of hidden layers	5	20	20
Number of units in layers	2	2	2
Weight of newer samples	1.0	2.5	2.5
Threshold for "new"	0.0	0.75	0.25
Loss function	MAE	MAE	MAE
Trend growth type	Continuous*	Continuous*	Discontinuous
Number of trend change-points	10*	10*	13
Fourier order for yearly seasonality	Auto*	Auto*	6
AR order	0*	6	6

4.3 Predictions of monthly averages

Results of predicting the average of calendar months can be seen in figures 13 through 16. It can be noted that all NP models perform worse than MHB during the months of July and December (except the exogenous Twitter model, which

cannot be tested this month), while performing on a similar level during August through October, which closely resemble corresponding months previous years.

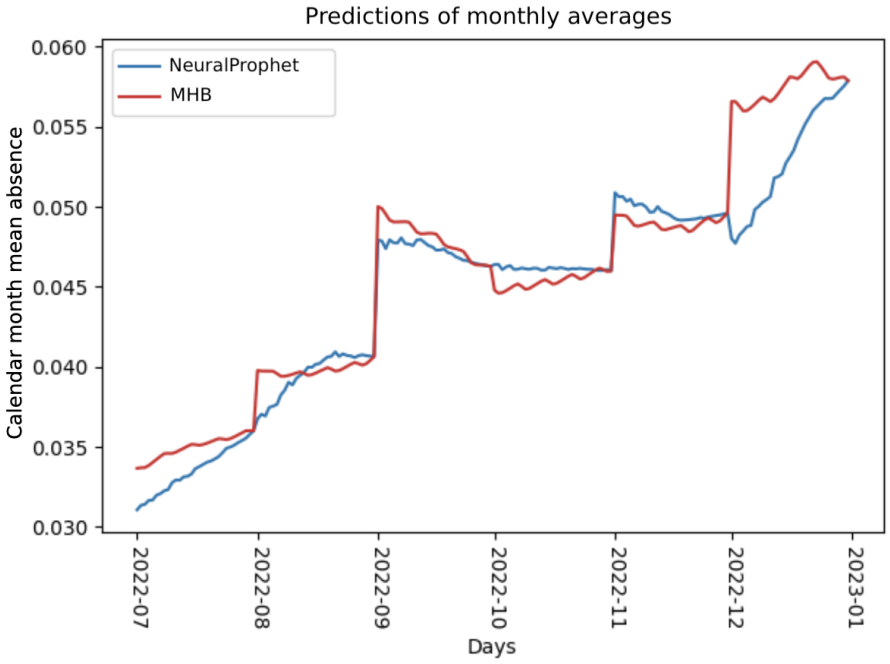

Figure 13: MHB vs. the univariate NeuralProphet model: Predictions of calendar month averages. The prediction is an average of predictions of the remaining days in the calendar month. A new prediction of the average is made each day and converge to the true average as the amount of measured data grows while the number of predictions is decreased.

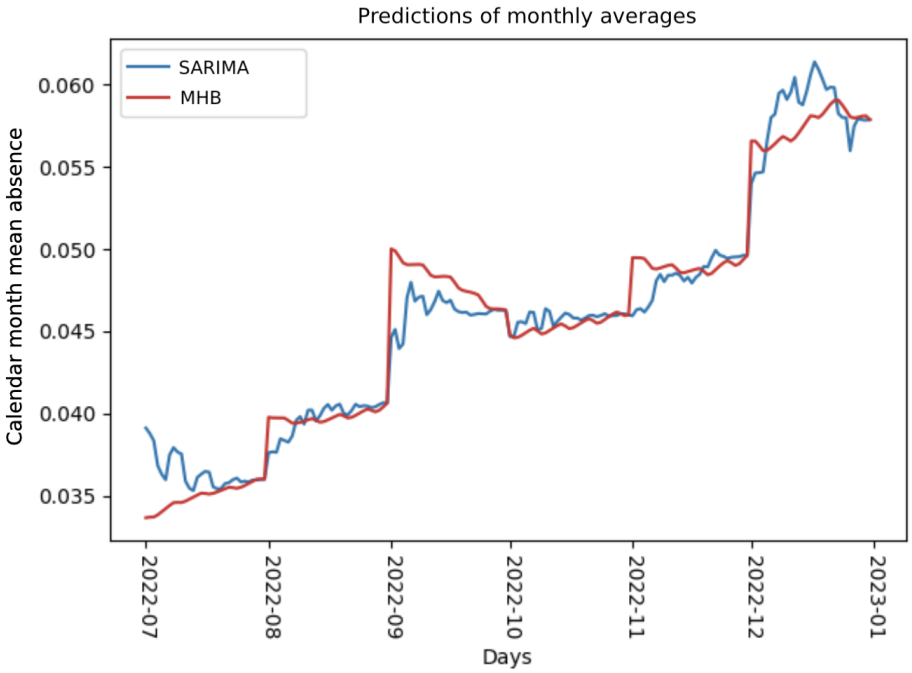

Figure 14: MHB vs. the SARIMA model: Predictions of calendar month averages. The prediction is an average of predictions of the remaining days in the calendar month. A new prediction of the average is made each day and converge on the true average as the amount of measured data grows while the number of predictions is decreased.

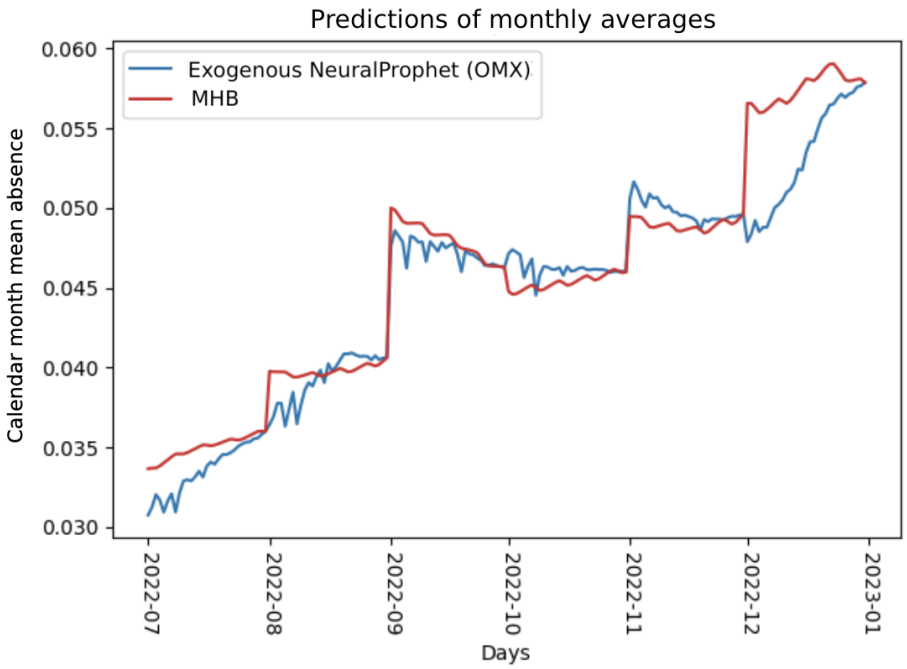

Figure 15: Baseline vs. the exogenous OMX NeuralProphet model: Predictions of calendar month averages. The prediction is an average of predictions of the remaining days in the calendar month. A new prediction of the average is made each day and converge on the true average as the amount of measured data grows while the number of predictions is decreased.

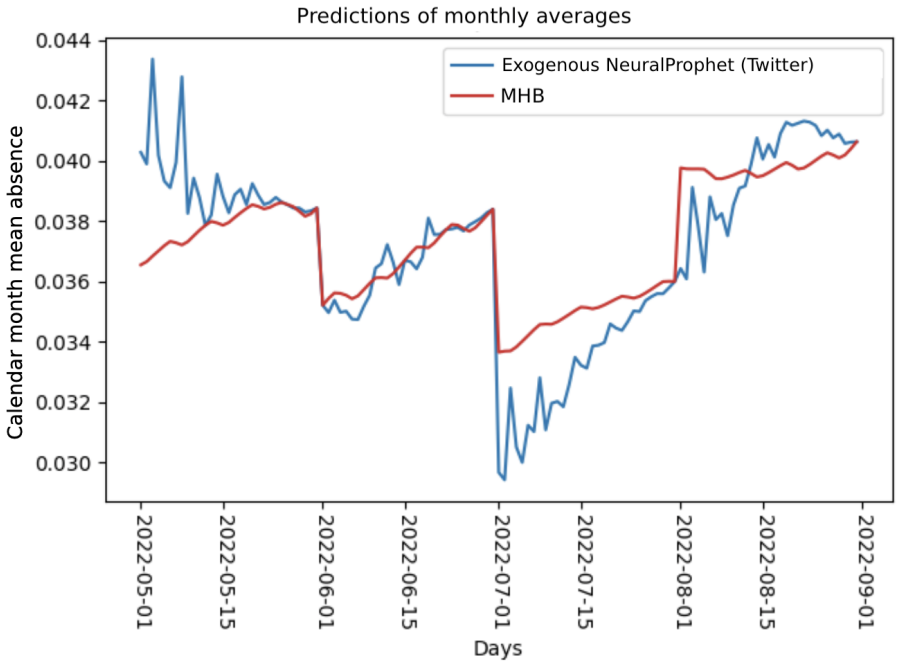

Figure 16: Baseline vs the exogenous Twitter NeuralProphet model: Predictions of calendar month averages. The prediction is an average of predictions of the remaining days in the calendar month. A new prediction of the average is made each day and converge on the true average as the amount of measured data grows while the number of predictions is decreased.

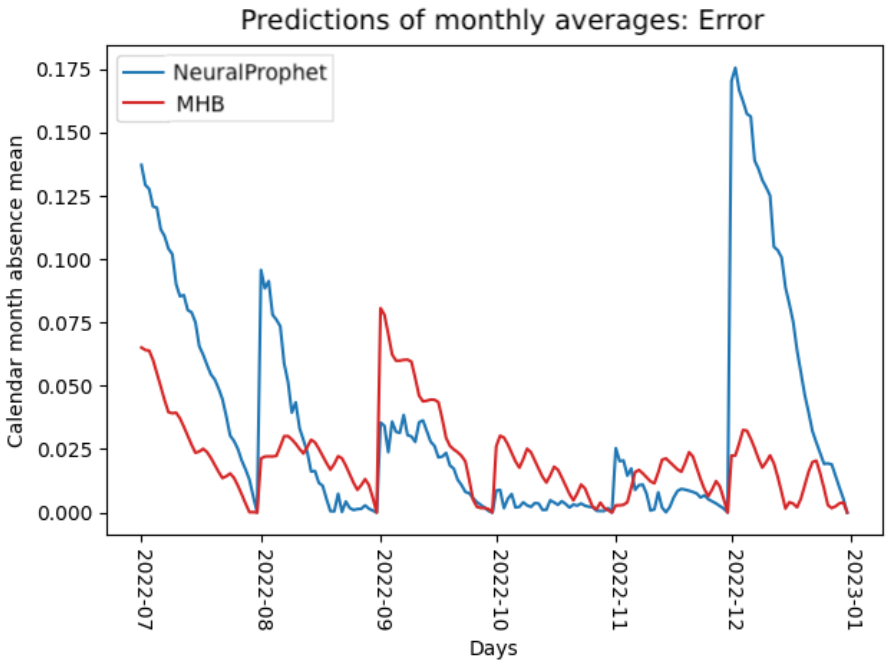

Figure 17: MHB vs. the univariate NeuralProphet model: Absolute errors of predictions of calendar month averages. As the number of the measured data points grows during the month, the prediction error approaches zero.

The corresponding error plots are shown in figures 17 through 20. It can be seen that the exogenous NP Twitter model performs far worse than any other models during July and August, the months where the test sets overlap. Furthermore, one can note the scale on the y-axis in figure 18 is different, illustrating the lower error. Additionally, it can be seen that there is no critical difference in prediction behaviour between the univariate and exogenous OMX NP model.

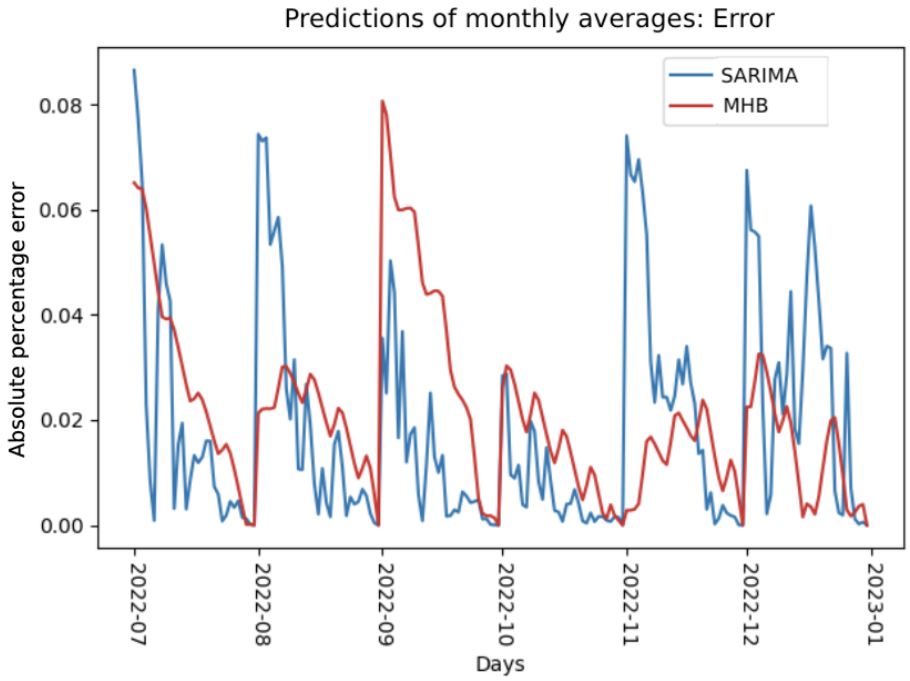

Figure 18: MHB vs. SARIMA model: Absolute errors of predictions of calendar month averages. As the number of the measured data points grows during the month, the prediction error approaches zero.

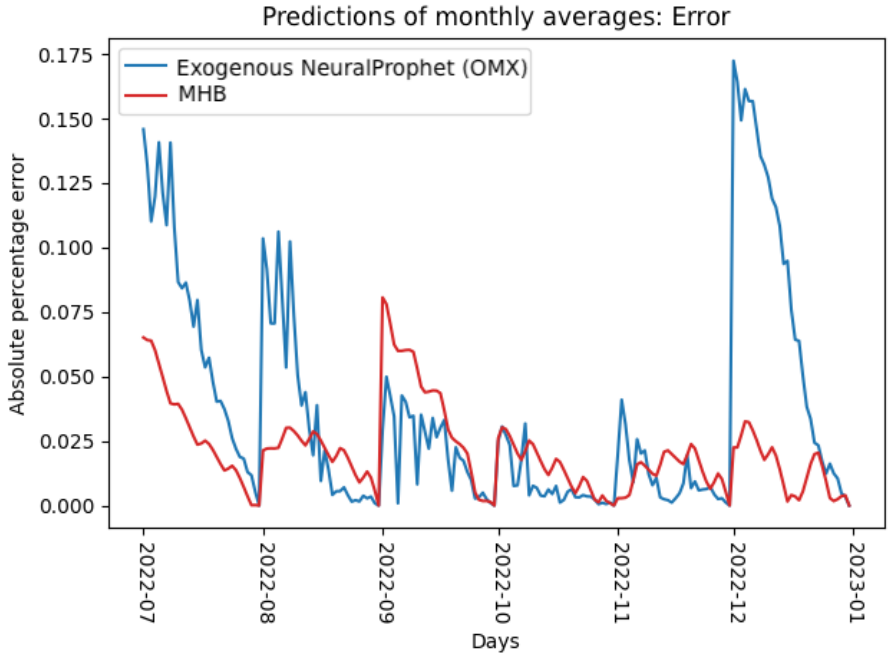

Figure 19: Baseline vs. the exogenous OMX NeuralProphet model: Absolute errors of predictions of calendar month averages. As the number of the measured data points grows during the month, the prediction error approaches zero.

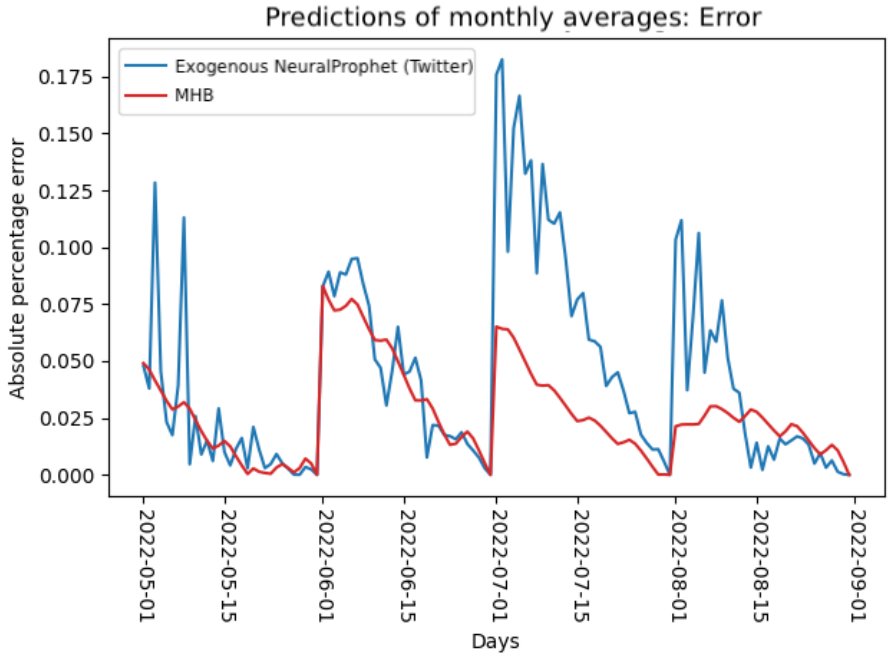

Figure 20: Baseline vs. the exogenous Twitter NeuralProphet model: Absolute errors of predictions of calendar month averages. As the number of the measured data points grows during the month, the prediction error approaches zero.

The resulting metrics of the calendar month predictions are presented in tables 3 through 5. It can be seen that the SARIMA model performs better than MHB concerning MAE and MAPE and while the MSE and variance are higher, the difference between the two metrics are lower, suggesting a lower bias in the SARIMA model.

Table 3: Metrics - Univariate models - Predictions of monthly averages, where predictions of averages are made on a daily basis, combining measured and predicted data.

Model	**MSE**	**Variance**	**MAE**	**MAPE**
NeuralProphet	$7.9 \cdot 10^{-6}$	$6.4 \cdot 10^{-6}$	$1.7 \cdot 10^{-3}$	$3.6 \cdot 10^{-2}$
SARIMA	$1.8 \cdot 10^{-6}$	$1.7 \cdot 10^{-6}$	$8.8 \cdot 10^{-4}$	$1.9 \cdot 10^{-2}$
MHB	$1.4 \cdot 10^{-6}$	$1.3 \cdot 10^{-6}$	$9.4 \cdot 10^{-4}$	$2.1 \cdot 10^{-2}$

Table 4: Metrics - Exogenous OMX - Predictions of monthly averages, where predictions of averages are made on a daily basis, combining measured and predicted data. AR coefficients have been used to incorporate the Swedish stock market into the predictions.

Model	**MSE**	**Variance**	**MAE**	**MAPE**
NeuralProphet	$8.0 \cdot 10^{-6}$	$6.6 \cdot 10^{-6}$	$1.7 \cdot 10^{-3}$	$3.7 \cdot 10^{-2}$
MHB	$1.4 \cdot 10^{-6}$	$1.3 \cdot 10^{-6}$	$9.4 \cdot 10^{-4}$	$2.1 \cdot 10^{-2}$

Table 5: Metrics - Exogenous Twitter - Predictions of monthly averages, where predictions of averages are made on a daily basis, combining measured and predicted data. AR coefficients have been used to incorporate the discourse regarding COVID-19 into the predictions.

Model	**MSE**	**Variance**	**MAE**	**MAPE**
NeuralProphet	$5.3 \cdot 10^{-6}$	$3.9 \cdot 10^{-6}$	$1.6 \cdot 10^{-3}$	$4.4 \cdot 10^{-2}$
MHB	$1.7 \cdot 10^{-6}$	$6.4 \cdot 10^{-7}$	$1.0 \cdot 10^{-3}$	$2.7 \cdot 10^{-2}$

4.4 Predictions of daily absence

Predictions of $h = 7$ with univariate models can be seen in figures 21 and 22. While the NP model tracks the slower season well, the SARIMA model also tracks the fine dynamics and is able to predict well during the peak in December 2022.

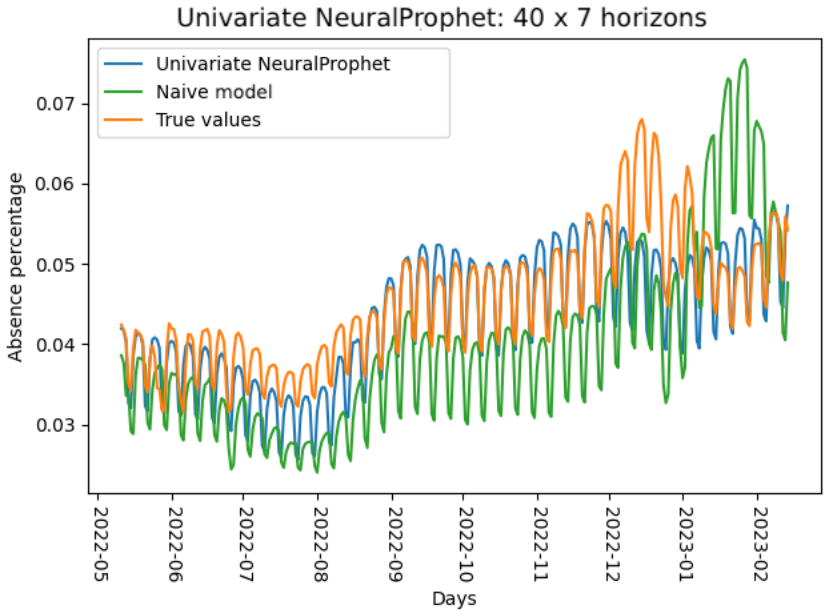

Figure 21: Univariate predictions using NeuralProphet: Predictions of daily sick absence with horizon length $h = 7$. This procedure is repeated until predictions have been formed for the whole test set.

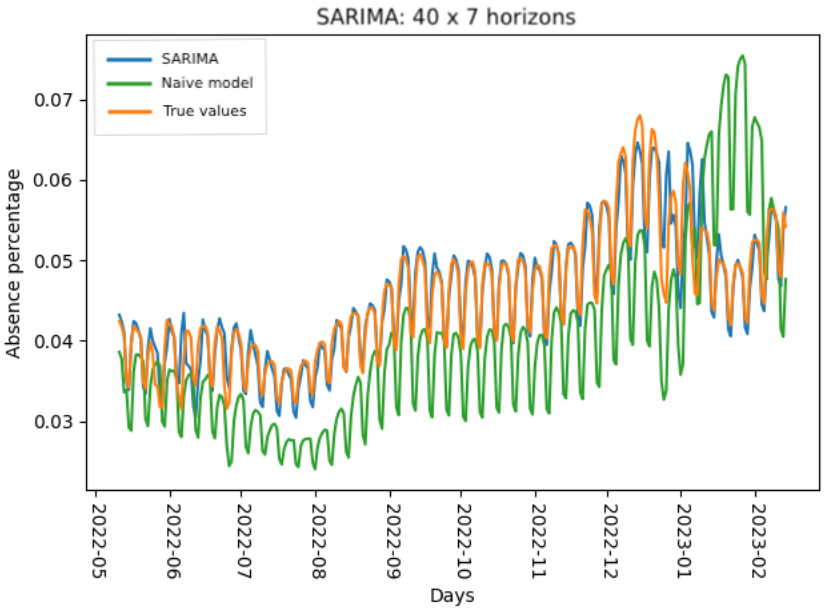

Figure 22: Univariate predictions using SARIMA: Predictions of daily sick absence with horizon length $h = 7$. This procedure is repeated until predictions have been formed for the whole test set.

Predictions of $h = 30$ with univariate models can be seen in figure 23 and 24. The NP model performs quite similar, but over- and undershoots slightly more. The SARIMA model displays a rather piece-wise behaviour per horizon, which results in worse predictions in comparison to $h = 7$.

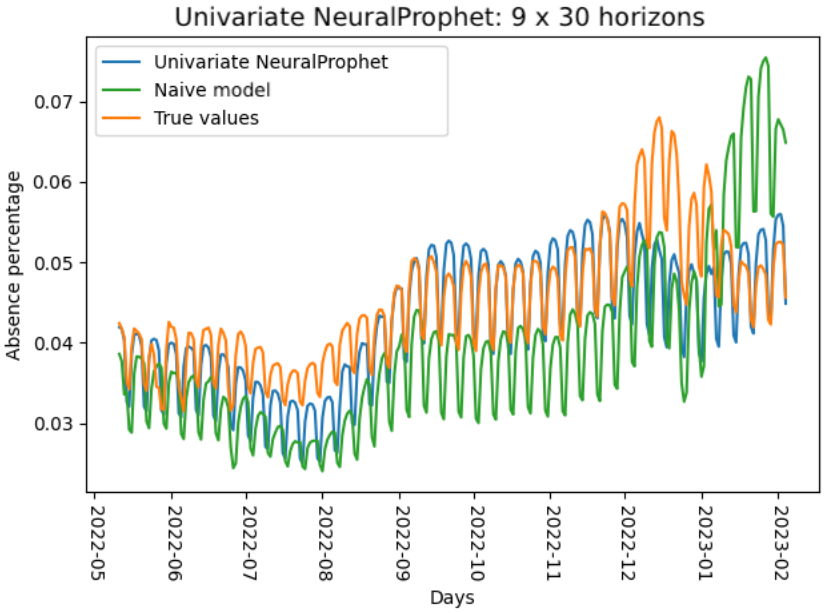

Figure 23: Univariate predictions using NeuralProphet: Predictions of daily sick absence with horizon length $h = 30$. This procedure is repeated until predictions have been formed for the whole test set.

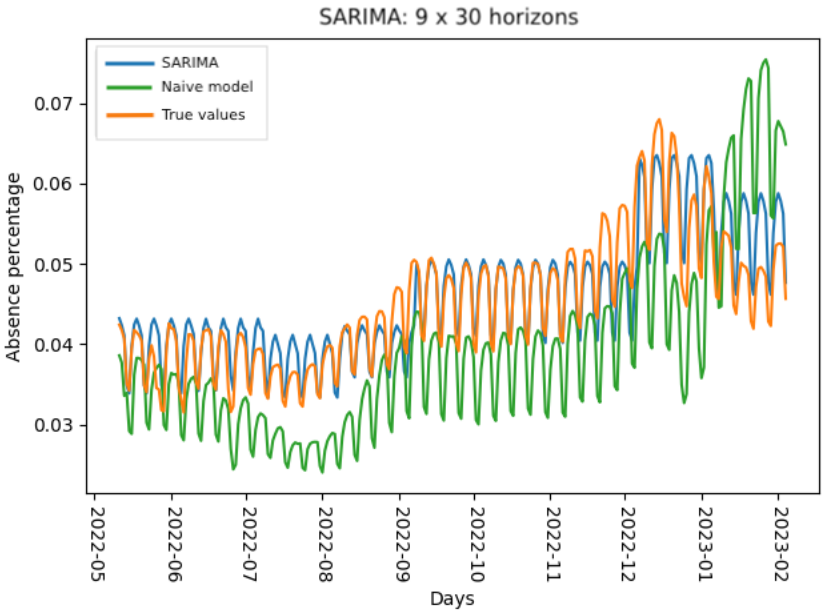

Figure 24: Univariate predictions using SARIMA: Predictions of daily sick absence with horizon length $h = 30$. This procedure is repeated until predictions have been formed for the whole test set.

Predictions of $h = 90$ with univariate models can be seen in figure 25 and 26. The piece-wise behavior of the SARIMA model is even more pronounced, while the NeuralProphet model predicts slightly lower values throughout most of the test set.

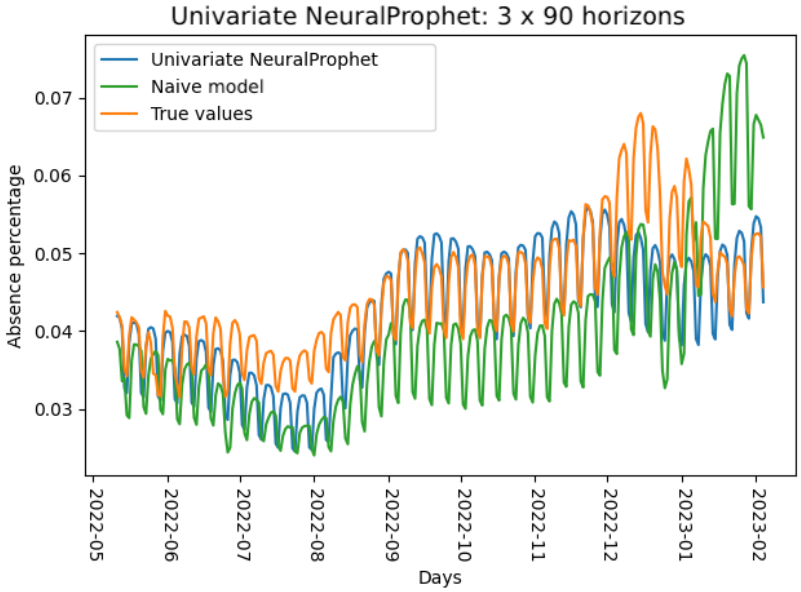

Figure 25: Univariate predictions using NeuralProphet: Predictions of daily sick absence with horizon length $h = 90$. This procedure is repeated until predictions have been formed for the whole test set.

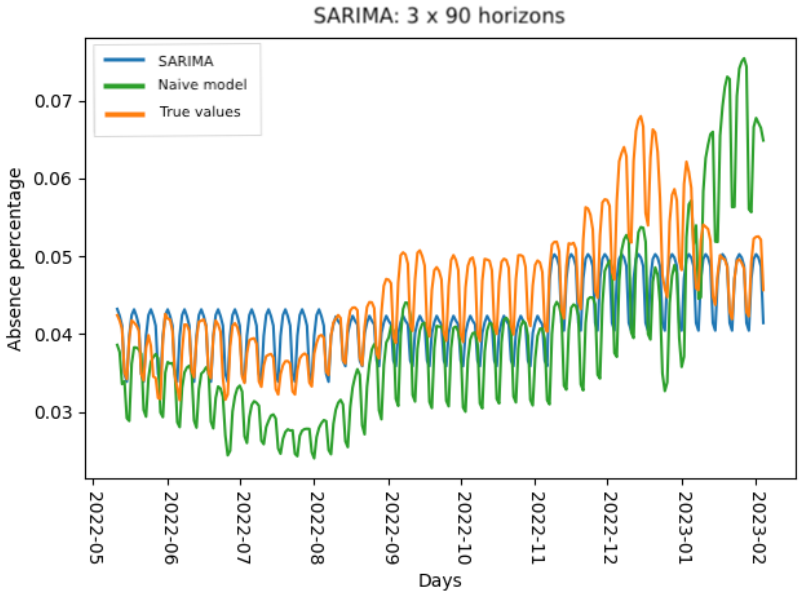

Figure 26: Univariate predictions using SARIMA: Predictions of daily sick absence with horizon length $h = 90$. This procedure is repeated until predictions have been formed for the whole test set

Corresponding metrics of all horizon lengths for the univariate models can be seen in tables 6 through 8. One can note that the best model regarding all metrics is the SARIMA model with $h = 7$ and that all models achieves at least half of the naive's metrics.

Table 6: Metrics - Univariate models - Predictions are of length $h = 7$, repeated 40 times in the test set and serve as the basis for the metrics. The naive model predicts the sick absence as identical to the previous year.

Model	**MSE**	**Variance**	**MAE**	**MAPE**
NeuralProphet	$2.2 \cdot 10^{-5}$	$1.8 \cdot 10^{-5}$	$3.4 \cdot 10^{-3}$	$7.3 \cdot 10^{-2}$
SARIMA	$6.3 \cdot 10^{-6}$	$6.3 \cdot 10^{-6}$	$1.6 \cdot 10^{-3}$	$3.5 \cdot 10^{-2}$
Naive	$1.1 \cdot 10^{-4}$	$8.0 \cdot 10^{-5}$	$9.2 \cdot 10^{-3}$	$2.0 \cdot 10^{-1}$

Table 7: Metrics - Univariate models - Predictions are of length $h = 30$, repeated 9 times in the test set and serve as the basis for the metrics. The naive model predicts the sick absence as identical to the previous year.

Model	**MSE**	**Variance**	**MAE**	**MAPE**
NeuralProphet	$2.6 \cdot 10^{-5}$	$2.0 \cdot 10^{-5}$	$3.7 \cdot 10^{-3}$	$8.1 \cdot 10^{-2}$
SARIMA	$1.2 \cdot 10^{-5}$	$1.2 \cdot 10^{-5}$	$2.5 \cdot 10^{-3}$	$5.4 \cdot 10^{-2}$
Naive	$1.2 \cdot 10^{-4}$	$8.2 \cdot 10^{-5}$	$9.3 \cdot 10^{-3}$	$2.0 \cdot 10^{-1}$

Table 8: Metrics - Univariate models - Predictions are of length $h = 90$, repeated 3 times in the test set and serve as the basis for the metrics. The naive model predicts the sick absence as identical to the previous year.

Model	**MSE**	**Variance**	**MAE**	**MAPE**
NeuralProphet	$2.7 \cdot 10^{-5}$	$2.1 \cdot 10^{-5}$	$3.8 \cdot 10^{-3}$	$8.3 \cdot 10^{-2}$
SARIMA	$3.5 \cdot 10^{-5}$	$2.6 \cdot 10^{-5}$	$4.5 \cdot 10^{-3}$	$9.2 \cdot 10^{-2}$
Naive	$1.2 \cdot 10^{-4}$	$8.2 \cdot 10^{-5}$	$9.3 \cdot 10^{-3}$	$2.0 \cdot 10^{-1}$

4.5 Exogenous

Crosscorrelation between input and output data are shown in figures 27 and 28. Both exogenous sets show significant correlation in relation to the sick absence.

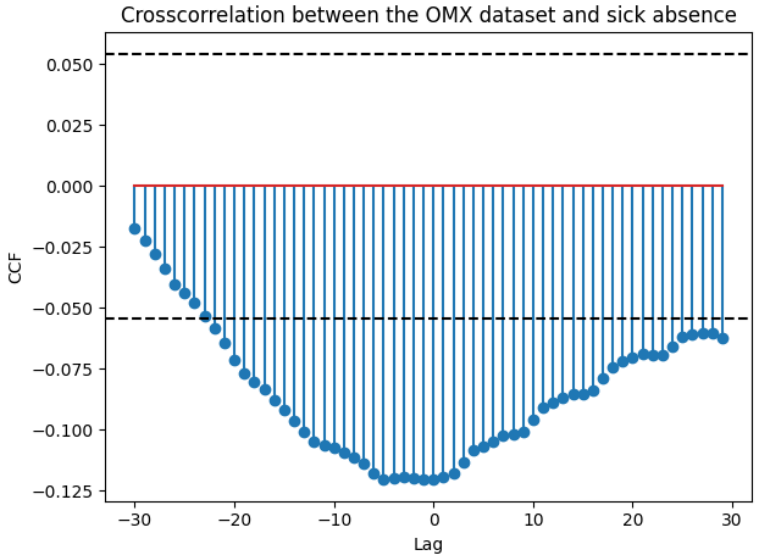

Figure 27: The crosscorrelation between the Swedish stock market (OMX30) and the sick absence showing their relationship over time. The sets are then fed to the NeuralProphet model for training. Note that the crosscorrelation is calculated using the modeling sets.

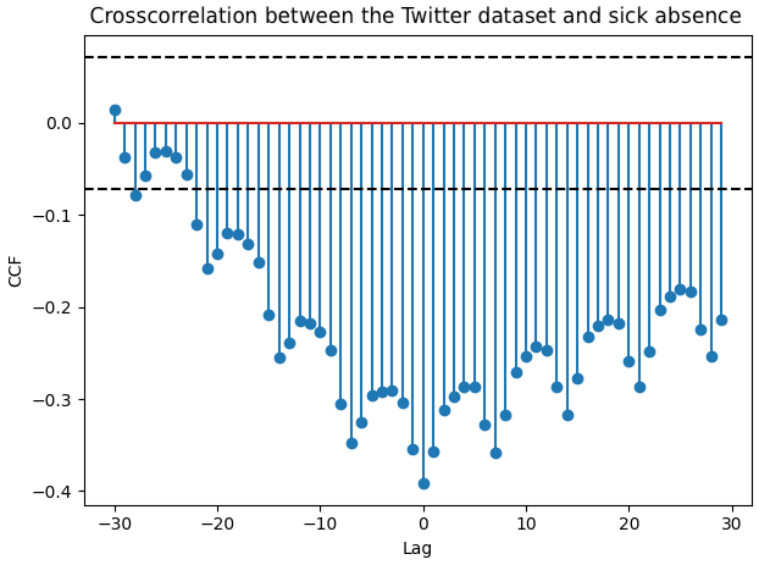

Figure 28: The crosscorrelation between the tweets regarding COVID-19 and the sick absence showing their relationship over time. The sets are then fed to the NeuralProphet model for training. Note that the crosscorrelation is calculated using the modeling sets.

Predictions for $h = 7$, $h = 30$ and $h = 90$ with the exogenous OMX model can be seen in figures 29, 30 and 31. For $h = 7$, the exogenous model undershoots less than its univariate counterpart. For $h = 30$ and especially $h = 90$, the baselines of the predictions are slightly lower the the univariate NP model.

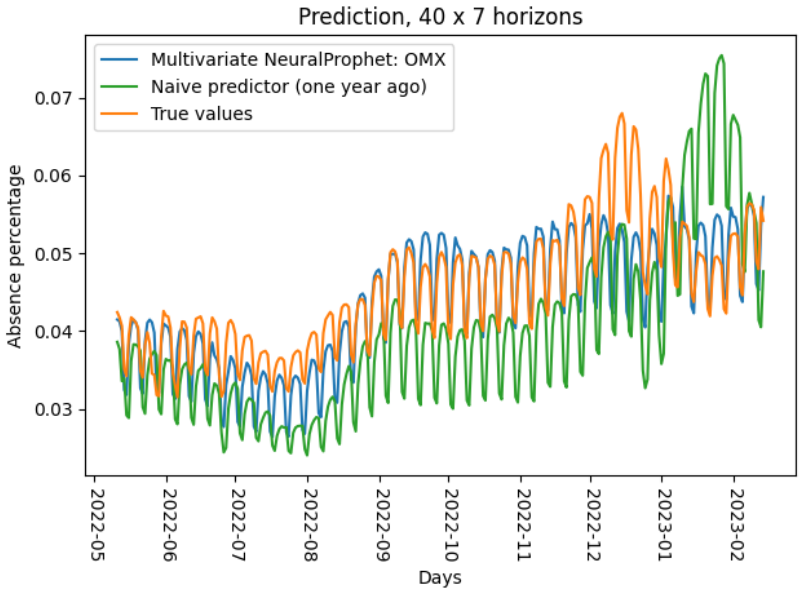

Figure 29: Predictions with exogenous input, OMX, that is incorporated with AR coefficients into NeuralProphet: Predictions are of daily sick absence with horizon length $h = 7$. This procedure is repeated until predictions have been formed for the whole test set.

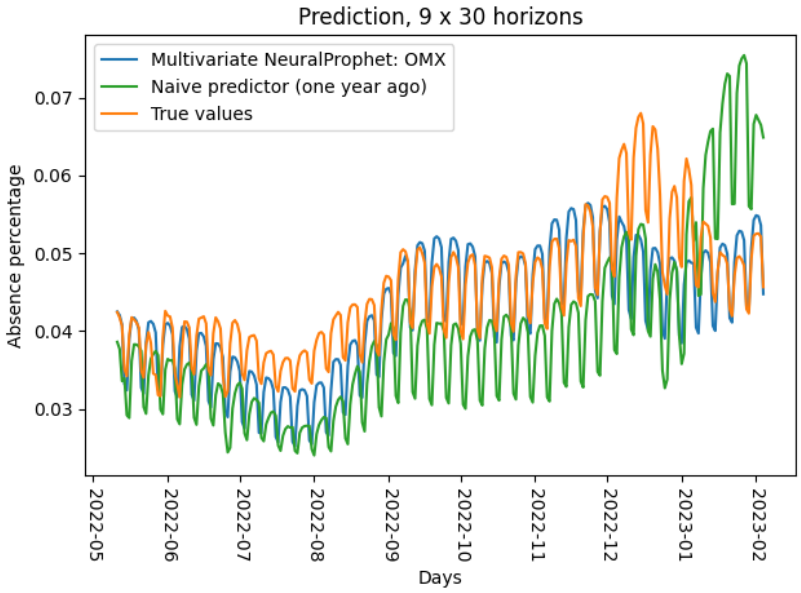

Figure 30: Predictions with exogenous input, OMX, that is incorporated with AR coefficients into NeuralProphet: Predictions are of daily sick absence with horizon length $h = 30$. This procedure is repeated until predictions have been formed for the whole test set.

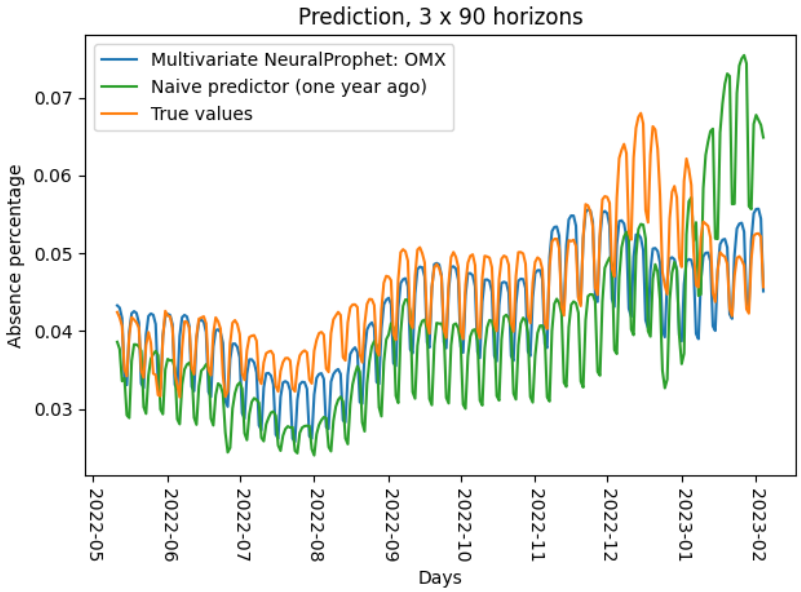

Figure 31: Predictions with exogenous input, OMX, that is incorporated with AR coefficients into NeuralProphet: Predictions are of daily sick absence with horizon length $h = 90$. This procedure is repeated until predictions have been formed for the whole test set.

The resulting metrics for the exogenous OMX model can be seen in table 9. For $h = 7$, the exogenous model performs better, achieving lower metrics overall. For the longer horizons, the metrics are very similar to the univariate counterpart.

Table 9: Metrics - Exogenous OMX model - Predictions are of length $h = 7$, 30 and 90 respectively. The naive model predicts the sick absence as identical to the previous year.

Model	**MSE**	**Variance**	**MAE**	**MAPE**
40 x 7	$1.9 \cdot 10^{-5}$	$1.6 \cdot 10^{-5}$	$3.1 \cdot 10^{-3}$	$6.9 \cdot 10^{-2}$
Corresponding naive	$1.1 \cdot 10^{-4}$	$8.0 \cdot 10^{-5}$	$9.2 \cdot 10^{-3}$	$2.0 \cdot 10^{-1}$
9 x 30	$2.6 \cdot 10^{-5}$	$1.9 \cdot 10^{-5}$	$3.7 \cdot 10^{-3}$	$8.1 \cdot 10^{-2}$
Corresponding naive	$1.2 \cdot 10^{-4}$	$8.2 \cdot 10^{-5}$	$9.3 \cdot 10^{-3}$	$2.0 \cdot 10^{-1}$
3 x 90	$2.6 \cdot 10^{-5}$	$1.7 \cdot 10^{-5}$	$3.8 \cdot 10^{-3}$	$8.4 \cdot 10^{-2}$
Corresponding naive	$1.2 \cdot 10^{-4}$	$8.2 \cdot 10^{-5}$	$9.3 \cdot 10^{-3}$	$2.0 \cdot 10^{-1}$

Predictions for $h = 7$, $h = 30$ and $h = 90$ with the exogenous Twitter model can be seen in figure 32 through 34. One can note that this model predict rather similarly to the naive model, i.e., the previous year.

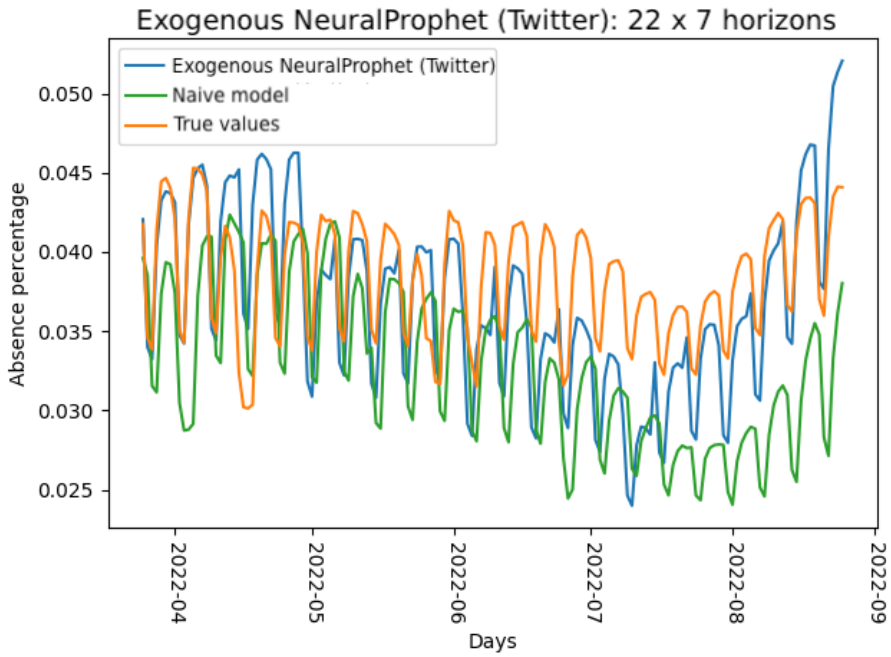

Figure 32: Predictions with exogenous input, tweets regarding COVID-19, that is incorporated with AR coefficients into NeuralProphet: Predictions are of daily sick absence with horizon length $h = 7$. This procedure is repeated until predictions have been formed for the whole test set.

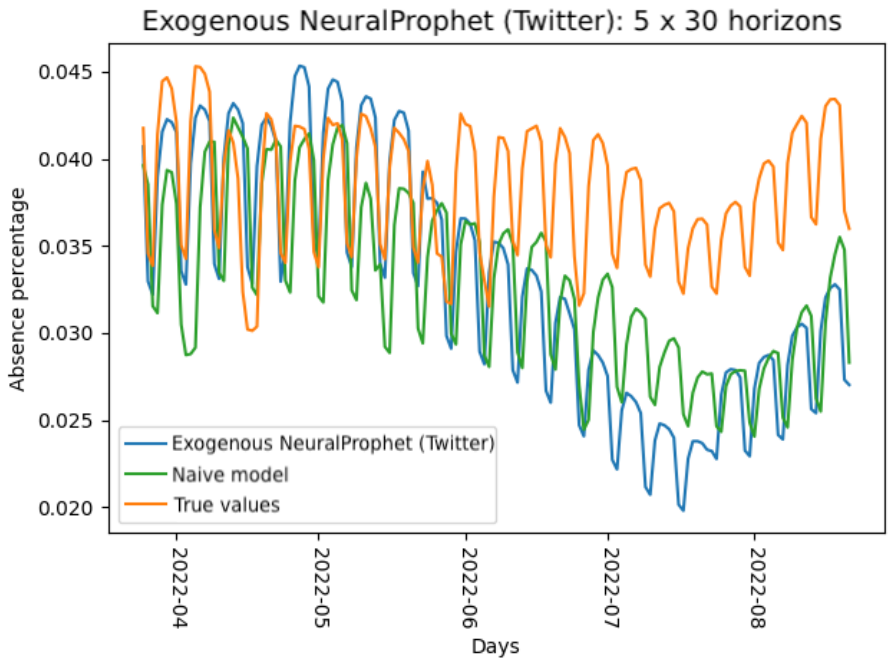

Figure 33: Predictions with exogenous input, tweets, that is incorporated with AR coefficients into NeuralProphet: Predictions are of daily sick absence with horizon length $h = 30$. This procedure is repeated until predictions have been formed for the whole test set.

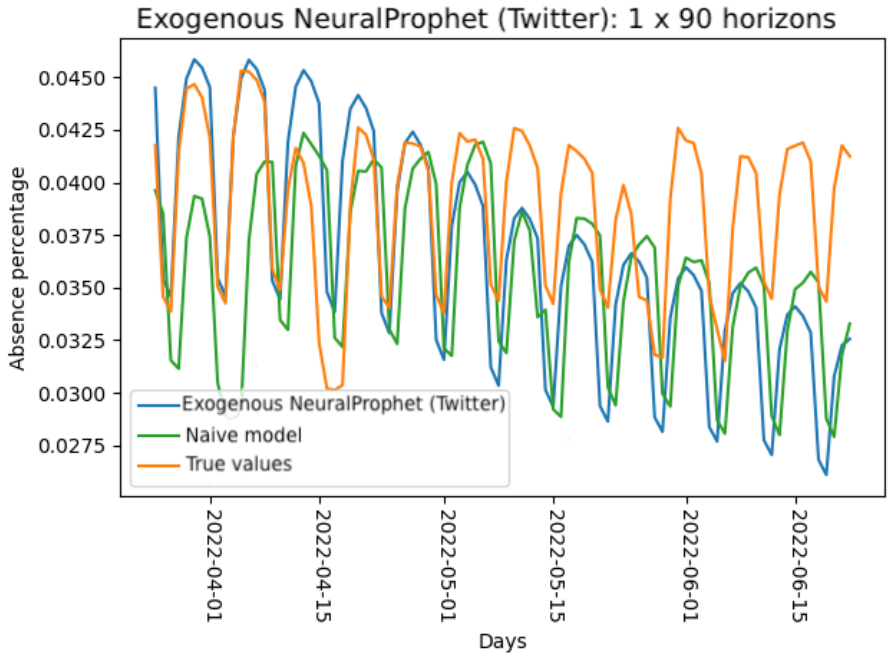

Figure 34: Predictions with exogenous input, tweets regarding COVID-19, that is incorporated with AR coefficients into NeuralProphet: Predictions are of daily sick absence with horizon length $h = 90$. This procedure is repeated until predictions have been formed for the whole test set.

The resulting metrics for the exogenous Twitter model can be seen in table 10. While the test set is different from the other models, it can be noted that it predicts less accurately (higher MAPE) than the other models for all horizon lengths; to the extent of being matched by the naive model for $h = 30$.

Table 10: Metrics - Exogenous OMX model - Predictions are of length $h = 7$, 30 and 90 respectively. The naive model predicts the sick absence as identical to the previous year.

Model	**MSE**	**Variance**	**MAE**	**MAPE**
22 x 7	$1.9 \cdot 10^{-5}$	$1.6 \cdot 10^{-5}$	$3.5 \cdot 10^{-3}$	$9.4 \cdot 10^{-2}$
Corresponding naive	$5.5 \cdot 10^{-5}$	$2.5 \cdot 10^{-5}$	$6.4 \cdot 10^{-3}$	$1.7 \cdot 10^{-1}$
5 x 30	$6.2 \cdot 10^{-5}$	$3.3 \cdot 10^{-5}$	$6.4 \cdot 10^{-3}$	$1.7 \cdot 10^{-1}$
Corresponding naive	$5.4 \cdot 10^{-5}$	$2.5 \cdot 10^{-5}$	$6.4 \cdot 10^{-3}$	$1.7 \cdot 10^{-1}$
1 x 90	$2.2 \cdot 10^{-5}$	$1.7 \cdot 10^{-5}$	$3.8 \cdot 10^{-3}$	$9.9 \cdot 10^{-2}$
Corresponding naive	$3.3 \cdot 10^{-5}$	$2.3 \cdot 10^{-5}$	$4.7 \cdot 10^{-3}$	$1.2 \cdot 10^{-1}$

5 Discussion

5.1 Processing of data

The processing of data is an essential part of the project and the sick absence quotient was quite affected by it, as can be see in figure 12. As most of these entities were companies who only reported present employees but never any absent individuals, they were effectively diluting the absence quotient, hence their removal resulted in higher levels of absence, especially for dates in 2018 and 2019. A possible explanation for this is that companies have gotten more proficient at using the platform with time and in doing so, increased the fidelity of their sick absence. It is also worth noting that the topic of sick absence has become increasingly more prominent in latter years, as the COVID-19 pandemic changed many workplaces' view on and approach to addressing sick absence. Throughout the time period that is spanned by the set, many workplaces have introduced or encouraged the possibility of working from home, which would to a larger extent allow people to carry out their work from rather than report in as ill [40]. Furthermore, the impact of the pandemic on the data is very clear, see the two first peaks in figure 12). The way in which anomalies (for this project, massive spikes in sick absence) were processed could have been handled in alternate ways. Rather that replacing them with the threshold value (+3 STD), the possibility of interpolating data points with a Kalman filter was discussed. However, the added amount of work required for such an action was deemed superfluous, as the total number of outliers, 28, was very limited. That being said, absences in periods surrounding these outliers (the pandemic outbreak, omicron) are still higher than corresponding periods of normal years. As the purpose of the thesis was to predict absence in "normal periods", one could argue that deleting COVID-19 from the set using data points from periods from previous years could result in models that are more apt for predicting normalcy, the presence of COVID-19 in the dataset was deemed too substantial.

Although the MedHelp dataset is very detailed and expresses more than 200.000 users, it is again worth noting that while it would be ideal to process, model and predict actual sick absence, the dataset is but a reflection of the *reported* absence. The results should therefore be treated as such. Moreover, there are other peculiarities in the data, which proves to affect model performance, December 2022 being one such example. This peak is not likely to have been caused solely by COVID-19, but rather by a mixture of tract respiratory infections [41].

5.2 Datasets

5.2.1 OMX

The OMX set used for modeling is negatively correlated with the sick absence, which would corroborate earlier findings of stock market impact on sick absence. OMX was chosen a measure of economy on a macro scale, but when considering which economic factors that impact households and affect behaviour more directly, it would have been interesting to instead measure consumer price index, CPI. Examining the sick absence through a lens of the development of living costs might have garnered even more promising results, however, CPI is only measured on a monthly basis, rendering its compatibility with a daily time series

such as sick absence subpar. However, one could construct a model where the CPI is used as a covariate with a constant value each calendar month and evaluate the performance of such a model. Other similar measures of economy suffer from similar problems. It should also be noted that despite the large proportion of Swedish inhabitants investing in the Swedish market, their investments are by no means limited to the domestic market and developments in direct development in foreign markets are not taken into account in this project. In other words, the OMX set is by no means exhaustive.

On the other hand, the length of the OMX set renders it otherwise compatible with the sick absence. More advanced methods for interpolating missing values could have an impact on training and consequent prediction accuracy, yet it is hard to motivate any particular interpolation method and ergo, the simple approach used for this project was deemed sufficient. Future work could explore how advanced interpolation methods could more accurately represent a time series such as stock markets.

5.2.2 Twitter

The Twitter set also exhibits a negative crosscorrelation with the MedHelp dataset. This is due to the inverse square root transform of the sick absence, which effectively mirrors the correlation with respect to the y-axis. While the TweetsCOV19 set holds the potential of exploring the connection between social discourse and sick absence (primarily COVID-19), there are some inherent problems that arise when using it. It only uses English tweets, which poses an issue as the pandemic was hardly limited to the English speaking part of the world. Although, Swedish people are generally prolific in the English language, it is hard to argue that the Swedish population would be represented by a dataset containing only information in English and that is not simultaneously not in any way limited to the Swedish population.

Moreover, the data length of the Twitter set constitutes another drawback. While it is only natural that discourse around the COVID-19 pandemic is not measured much earlier than the pandemic outbreak, its limit in scope requires removing more than 1.5 years of data (January 2018 - September 2019, arguably the most "normal" part) from the beginning of the set, and 5 months at the end.

5.3 Univariate models

As was expected, predictions of shorter horizon lengths have lower MAPE, which can be seen in table 6. The NP performance is worsened when increasing the length of the horizons, however, the difference in performance is not vastly different between 7 and 30 and 30 and 90, despite the large difference in added guesses (23 compared 60). There is no clear reason to why the performance is not reduced to an even greater degree, but it could be argued that the seasonality of the data contributes to the somewhat maintained accuracy. The yearly seasonality, in particular, informs NeuralProphet predictions and given a "normal" year, this will imply adequate performance. However, during abnormal periods, the seasonality impair this model more than it aids. As an example, one may study the month of December 2022: the absence during this month is

higher than its earlier counterparts and as a consequence, this period hosts the worst predictions of the NeuralProphet model.

The SARIMA model, however, takes the recent lags into consideration and does not suffer to the same extent and is able to track the dynamics very well. Its model predictions for $h = 7$ has the lowest metrics of all and exhibits no noteworthy tendency for errors. These predictions are able to track both slow and fast dynamics, while simultaneously maintaining the correct amplitude. The MAE, being within a range of one tenth of a percent, could inform business and decision makers on how sick absence will vary over the week with good resolution and in doing so facilitate staffing. For $h = 30$, the results are worse but the offset of the predictions are still similar enough to the true values to generate low metrics. With the growing prediction horizon, the reduced visibility of yearly seasonality within the predictions is becoming more apparent. For $h = 90$, the predictions in the test set essentially form a piece-wise function with three distinct sub-signals with a respective frequency and offset. For $h = 30$ and 90, the horizons are simply not sufficiently stationary, and the resulting predictions become very unanimous, which cause an increase in the metrics. The insistence on seasonality in the NP model generate more accurate predictions for this horizon and the advantages and drawbacks of the model types are exemplified through their different emphasis on stationary data and observed seasonality.

The main reason for the discrepancy between the true data and naive predictions is mostly due to an offset rather than a matter of faulty dynamics. The only major difference in behaviour occurs in late January, when the naive model echoes the effects of the Omicron variant of COVID-19, which caused a surge in Swedish sick absence. One could use a naive which forms its prediction for a given calendar day by averaging earlier instances of that calendar day, but that poses two problems. Firstly, it eliminates the simple nature of a naive model and secondly generates a worse performing naive as the sick absence during earlier years was generally lower.

5.4 Exogenous models

As can be seen in the metrics from the exogenous models (see table 9 and 10), results from using external input vary. Both exogenous models have a higher number of tweaked hyperparameters than the univariate model counterpart, and model complexity can therefore not be necessarily accredited to being paramount for performance. It is rather the nature of the exogenous dataset and its inherent connection to the sick absence that appears to affect results.

When considering a short forecast horizon, using previous AR lags in combination with seasonal terms proved successful when incorporating the OMX set into a NeuralProphet model, as a lower MAPE is achieved. This proves in line with the previously mentioned research that found a connection between the stock market and sick absence [1]. It is worth noting that there is no way of comparing the severity of the sick absences in this project and that, but the increased accuracy of the exogenous OMX model could lend credence to the thesis that stock markets indeed influence sick absence. Further speculation on the matter is more apt for a project focused on causality. Whatever the nature of the relationship between the OMX set and the MedHelp dataset is,

longer prediction horizons are mostly unaffected when considering the metrics. The information in the six last OMX datapoints fail to provide any information relevant enough to firmly improve long-term predictions. Most metrics are very similar except the variance, which is lower for the exogenous model. The fact that this metric differs could be attributed to the fact that the exogenous OMX model has an reduced amplitude in the yearly seasonality during the period of the test set. This in turn might be a result of the stock market working as a stabilizing agent in relation to the sick absence as the usage of OMX decreases seasonality over and undershooting. Moreover, the impact of more recent observations should also be taken into account. Despite initial concerns that increasing the weight of newer samples would cause the NP models to incorrectly learn from COVID-19 and consequently form inferior predictions, the exogenous OMX model dismisses this thesis. The emphasis on more recent observations and inclusion of AR-components in combination with exogenous OMX data works in tandem to create more accurate short-term predictions for this test set. A reason as to why the long-term accuracy is largely unaffected could be the way in which the exogenous input are handled in NeuralProphet. If the input would be modeled in a similar manner as the absence, i.e, with trends and seasons rather than just autoregression, it is possible that the long-term predictions could be improved.

The results of when incorporating social discourse in the form om the TweetsCOV-19 set are not as promising. Although it would have been interesting to potentially discover that social media caused people to report absence, it is from the predictions results apparent that such a finding is unlikely. The Twitter data exhibits significant negative correlation with the sick absence, which could be explained by the inverse square root transform used for the this sick absence set. The weekly seasonality is very apparent; possibly a result of a lag when reporting in COVID-19 cases. However, the inadequate data length in combination result in metrics worse than the univariate NP model. A reason for the worsened performance when incorporating the Twitter could also be that the information supplied external input is not sufficient to compensate for the increased complexity of the model. The difference in complexity can be seen in the parameters in in table 2, e.g., through a more complex neural network. It is also possible that the tweets and the sick absence are too synchronized for the tweets to provide enough useful information and in doing so, increase prediction accuracy.

It would be interesting to further evaluate the impact of using the Twitter set without concerns regarding if there is enough sufficient historical data, however, considering the current datasets this is not feasible. In order to generalize the results and draw a well-founded verdict regarding the incorporation of exogenous macro factors, further testing when more data is available is required.

Morever, exploring the macro factors impact with SARIMA(X) or Box-Jenkins models and investigate if even better results could be achieved, would be of interest and a natural next step in this field of research when adequate tools are available.

5.5 Predictions of calendar monthly averages

The univariate NP model surpasses MHB September through November in terms of prediction accuracy (MAPE), while performing worse July, August and December, see figure 17. A possible reason for this is that September through November (while on a generally higher level, i.e., offset) behave very similar to previous corresponding months of earlier years. As the NP models are retrained on a daily basis, it is allowed to adjust the trend in the data up until the last 20% of training data. The rising trend in the data that can be seen in, e.g., figure 5 can therefore be captured, which could explain why the NP model handle the elevated level of absence. On the other hand, for short-lasting variations in the data, it performs worse. The regular dip in absence in July is not as pronounced as in previous years, which could explain why it undershoots during this month. For November and particularly December, patterns of absence differ greatly from corresponding periods during the years 2018-2020. The surge in absence levels, however, is to a lesser extent present in 2021. As the initial MHB guess of the month is very much dictated by the previous year and moth, it does not undershoot nearly as much as the univariate NP model when considering the month of December.

The SARIMA model outperforms MHB slightly, see figure 14 as well as table 3, and exhibits a more turbulent behaviour that its NP counterpart. A reason for this could be the increased consideration of recent lags affecting each new set of predictions. While the MAPE is quite high during the first days of a calendar month, it generally declines quite quickly. It should also be noted that figure 17 and 18 have different scales on their respective y-axes. With this taken into account, the contrasting regarding performance becomes more evident.

The exogenous OMX NP model performs similarly to its univariate counterpart, but also displays a more erratic nature. The added consideration of capturing more fast-changing variations using AR-components could be attributed to this. A similar behaviour can be seen in 16, but the resulting metrics are quite different. It should be restated that the Twitter has a different test set and can not be directly compared to the others, but inspecting each models respective metrics can be done to extract some insight into performance. While the other metrics suggest better performance, see table 5, the MAPE makes it apparent that is not the case. The average level of absence during these months are lower (which can be seen by comparing the y-axes in figures 15 and 16). As the MAPE reduces the impact of scale on the data, the Twitter NP model is shown to have the largest percentage error out of all models.

5.6 Complexity

Implementing the prediction methodology for the different types of models is a complicated problem. The implementation was considerably more time-consuming than predicted but the algorithms could all the same be optimized for performance. For this project, however, these matters of time complexity was secondary to a solution that simply generated the intended result. Another aspect of time complexity to consider is the usage of grid searches. One could theoretically ensure utilize an infinite search space for the objective of finding optimal hyperparameters for NP models, but in practice, there are issues that

arise. Primarily, conducting grid searches is very computationally demanding and by extension, time consuming. There is the possibility of outsourcing the computations to a GPU, using on-premise equipment or web services, but resources of this nature were not available for this project. Nevertheless, the search space could have been expanded using randomized grid searches, potentially improving model prediction accuracy.

While using NeuralProphet appears to generate more accurate predictions and higher performing models for long horizons ($h = 90$), in practice, there are other factors that would inform the selection of model type. For models on this scale, the training and and prediction time for a machine learning model such as NeuralProphet is tenfold for regular prediction horizons and almost thousandfold when forming the predictions for the calendar month test set. The number of batch sizes and epochs could of course be tweaked to decrease training times, but at the cost of accuracy. If models for predicting daily sick absences were to be deployed for, e.g., each individual company using the MedHelp platform, it is likely that the difference in computational demand between SARIMA and NP would become very apparent. On the other hand, SARIMA models generally require more manual tweaking of matters such as polynomial orders, which also plays a role when considering matters of scalability and the feasibility of developing individually tailored models. An ideal type of model would be both highly automated regarding model parameter selection in addition to being very quick when it comes to training and predicting.

Further exploring matters of complexity and interpretability, it could be argued that the nature of MHB, albeit rudimentary, produces quite good results with essentially zero preprocessing required. The advantages of a "less is more" argument are apparent, see, e.g., table 3 as it performs better than any NP model, but it is worth noting also that the task of predicting calendar month average absences is not the priority of neither NeuralProphet nor SARIMA. While the average of calendar month very much is a result of daily predictions, the secondary nature of such results should be taken into account: MHB is designed specifically to do one task: predict the calendar month average, it is anticipated to be on par with models such as NP and SARIMA that are designed for another task (predictions with a daily resolution), despite its simple nature. If calendar month average absences had been the main focus of this thesis, the data could have been preprocessed in an alternate manner: Aggregating the data to a monthly basis, rather than daily, and then proceeding with the modeling in a similar way. However, in doing so, the number of datapoints available for modeling would be greatly reduced (by a factor of approximately 30), which in turn is very likely to affect performance results in a negative manner. It would also greatly reduce the number of predictions to 6, which is not sufficient for drawing any conclusion using statistical measures. On the other hand, given enough historic data, modeling calendar month absences is very much a potential project to explore.

However, there are aspects of the NeuralProphet models that provide more insight into the data than its counterparts. Its modular nature increases interpretability of the structure and variation of the data, such as the frequency components (e.g., yearly and weekly seasonalities) and general trends in the data. Despite this, the lack of a possibility to explicitly display the numeric

parameters of the model poses a minor black-box issue. The shortage of other prominent machine learning packages for time series analysis with a comparable performance and features available designates it the current best choice.

6 Conclusions

From this project we conclude that it is possible, through time series modeling, to predict daily sick absence with less than 9 % average error for weekly, monthly and quarterly horizons. Using a SARIMA model is better for $h = 7$ and 30, while NeuralProphet proves to be a superior choice for h 90. The difference in performance could be explained by contrasting emphases on autoregression and long-term seasonality. By processing the MedHelp data through removal of atypical companies, the dilation of the set is reduced, which increases the levels of sick absence.

Incorporating exogenous input requires more complex models and that the data input is related to the sick absence in order to compensate for this. Using OMX data increases prediction accuracy for weekly horizons, while Twitter data decreases performance for all horizon lengths. The length of the data sets used is also likely to affect performance and when more extensive data is available, the impact should be evaluated. The results are promising and implies possibilities to improve staffing measures, but would require further testing using other models to verify.

The complexity of the models infers a considerable amount of manual labour, especially concerning SARIMA. The computational power required for model training and optimization further complicates the issue. These matters decreases their feasibility to be deployed in real-world applications, where ideally, models could be bespoke on a company basis.

7 Ethical Aspects

7.1 Data

There are several aspects to consider concerning ethics within E-health. Among these are integrity, confidentiality and privacy [42].

The data collection necessary for MedHelp's platform strictly adheres to GDPR. Policys are in place to protect each individual data and their respective right to it. The data supplied to MedHelp through customers are based on consent and the same consent is also valid for research and analysis, e.g projects like this master thesis.

When considering the more practical elements, it is worth noting that the data supplied for this project neither features direct identifiers to the individuals (i.e. names or social security numbers) nor which companies actually constitute the dataset. The models are created using data represented as a population, meaning it is not possible to track specific individuals.

The nature of which the dataset is interacted with is one-way. In practice, this means that whichever way the data is processed, it will never affect the original, but merely a copy of it. Furthermore, the data was only allowed to be run in RAM for security reasons.

7.2 Intentions

This project serves as a part of MedHelp's goal of assisting companies manage and decrease work absence. It would be beneficial for companies to better prepare for and handle absence and also learn what might factors might drive the data, rather than trying to target specific individuals who might be skipping work.

References

[1] Joseph Engelberg and Christopher A. Parsons. "Worrying about the Stock Market: Evidence from Hospital Admissions". In: *The Journal of Finance* 71.3 (2016), pp. 1227–1250. DOI: https://doi.org/10.1111/jofi.12386. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jofi.12386. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.12386.

[2] George Bojas. *Labour Economics,* 7th ed. McGraw Hill, 2015. ISBN: 9780078021886.

[3] Örjan Lutz, Gunnar Sundqvist, and Bodil Umegård. "Sjukfrånvaro i kommuner och landsting - Vad är problemet?" In: (2017).

[4] Andreas Jakobsson. *An Introduction to Time Series Modeling.* Studentlitteratur, 2019. ISBN: 9789144134031.

[5] Anna Broman and Gabrielle Larsson. "Patterns of absenteeism – different during major sporting events?" In: (2015).

[6] Hanna Hultin et al. "Short-term sick leave and future risk of sickness absence and unemployment - The impact of health status". In: *BMC public health* 12 (Oct. 2012), p. 861. DOI: 10.1186/1471-2458-12-861.

[7] M Marmot et al. "Sickness absence as a measure of health status and functioning: from the UK Whitehall II study." In: *Journal of Epidemiology & Community Health* 49.2 (1995), pp. 124–130. ISSN: 0143-005X. DOI: 10.1136/jech.49.2.124. eprint: https://jech.bmj.com/content/49/2/124.full.pdf. URL: https://jech.bmj.com/content/49/2/124.

[8] Kristin Byron and Suzanne Peterson. "The impact of a large-scale traumatic event on individual and organizational outcomes: exploring employee and company reactions to September 11, 2001". In: *Journal of Organizational Behavior* 23.8 (2002), pp. 895–910. DOI: https://doi.org/10.1002/job.176. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/job.176. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/job.176.

[9] Eric Patton and Gary Johns. "Context and the social representation of absenteeism: Absence in the popular press and in academic research". In: *Human Relations* 65.2 (2012), pp. 217–240. DOI: 10.1177/0018726711428819. eprint: https://doi.org/10.1177/0018726711428819. URL: https://doi.org/10.1177/0018726711428819.

[10] Richard R. Tansey and Michael R. Hyman. "Public Relations, Advocacy Ads, and the Campaign Against Absenteeism During World War II". In: *Business Professional Ethics Journal* 11.3/4 (1992), pp. 129–163. ISSN: 02772027. URL: http://www.jstor.org/stable/27800891 (visited on 05/09/2023).

[11] Axel Bruns and Katrin Weller. "Twitter as a First Draft of the Present: And the Challenges of Preserving It for the Future". In: WebSci '16. Hannover, Germany: Association for Computing Machinery, 2016, pp. 183–189. ISBN: 9781450342087. DOI: 10.1145/2908131.2908174. URL: https://doi.org/10.1145/2908131.2908174.

[12] Dimitar Dimitrov et al. “TweetsCOV19 - A Knowledge Base of Semantically Annotated Tweets about the COVID-19 Pandemic”. In: *Proceedings of the 29th ACM International Conference on Information Knowledge Management*. CIKM ’20. Virtual Event, Ireland: Association for Computing Machinery, 2020, pp. 2991–2998. ISBN: 9781450368599. DOI: 10.1145/3340531.3412765. URL: https://doi.org/10.1145/3340531.3412765.

[13] Emily Chen, Kristina Lerman, and Emilio Ferrara. “Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set”. In: *JMIR Public Health Surveill* 6.2 (May 2020), e19273. ISSN: 2369-2960. DOI: 10.2196/19273. URL: http://www.ncbi.nlm.nih.gov/pubmed/32427106.

[14] Paul Edwards and Kay Greasley. *Absence from work*. Accessed on May 9, 2023. URL: https://www.eurofound.europa.eu/publications/report/2010/absence-from-work.

[15] *Ersättning för karens under coronapandemin*. Accessed on Feb 10, 2023. URL: https://www.forsakringskassan.se.

[16] Brad Comincioli. “The Stock Market as a Leading Indicator: An Application of Granger Causality”. In: *The Park Place Economist* 1 (1996), p. 13.

[17] Ross Levine and Sara Zervos. “Stock Market Development and Long-Run Growth”. In: *The World Bank Economic Review* 10.2 (1996), pp. 323–339. ISSN: 02586770, 1564698X. URL: http://www.jstor.org/stable/3990065 (visited on 05/03/2023).

[18] *Euroclear Sweden: Aktieägandet i Sverige 2022*. Accessed on May 8, 2023. URL: https://www.euroclear.com/sweden/sv/det-svenska-aktieagandet.html.

[19] Angeliki Menegaki. “Chapter 2 - Stationarity and an alphabetical directory of unit roots often used in the energy-growth nexus”. In: *A Guide to Econometrics Methods for the Energy-Growth Nexus*. Ed. by Angeliki Menegaki. Academic Press, 2021, pp. 31–61. ISBN: 978-0-12-819039-5. DOI: https://doi.org/10.1016/B978-0-12-819039-5.00002-1. URL: https://www.sciencedirect.com/science/article/pii/B9780128190395000021.

[20] Rob J. Hyndman and George Athanasopoulos. *Forecasting: principles and practice*. Otexts, 2021. ISBN: 9780987507136.

[21] Skipper Seabold and Josef Perktold. “statsmodels: Econometric and statistical modeling with python”. In: *9th Python in Science Conference*. 2010.

[22] Adam Paszke et al. “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[23] *What is a neural network*. Accessed on May 17, 2023. URL: https://www.ibm.com/topics/neural-networks.

[24] Moritz Hardt, Ben Recht, and Yoram Singer. "Train faster, generalize better: Stability of stochastic gradient descent". In: *Proceedings of The 33rd International Conference on Machine Learning*. Ed. by Maria Florina Balcan and Kilian Q. Weinberger. Vol. 48. Proceedings of Machine Learning Research. New York, New York, USA: PMLR, 20–22 Jun 2016, pp. 1225–1234. URL: https://proceedings.mlr.press/v48/hardt16.html.

[25] Léon Bottou. "Stochastic Gradient Descent Tricks". In: *Neural Networks: Tricks of the Trade: Second Edition*. Ed. by Grégoire Montavon, Geneviève B. Orr, and Klaus-Robert Müller. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 421–436. ISBN: 978-3-642-35289-8. DOI: 10.1007/978-3-642-35289-8_25. URL: https://doi.org/10.1007/978-3-642-35289-8_25.

[26] Oskar Triebe, Nikolay Laptev, and Ram Rajagopal. "AR-Net: A simple Auto-Regressive Neural Network for time-series". In: *CoRR* abs/1911.12436 (2019). arXiv: 1911.12436. URL: http://arxiv.org/abs/1911.12436.

[27] Sean J. Taylor and Benjamin Letham. "Forecasting at Scale". In: *PeerJ Preprints* 5:e3190v2 (2017). DOI: 10.7287/peerj.preprints.3190v2.

[28] Oskar Triebe et al. "NeuralProphet: Explainable Forecasting at Scale". In: *CoRR* abs/2111.15397 (2021). arXiv: 2111.15397. URL: https://arxiv.org/abs/2111.15397.

[29] H. Akaike. "A new look at the statistical model identification". In: *IEEE Transactions on Automatic Control* 19.6 (1974), pp. 716–723. DOI: 10.1109/TAC.1974.1100705.

[30] Greta M. Ljung and G. E. P. Box. "On a measure of lack of fit in time series models". In: *Biometrika* 65 (1978), pp. 297–303.

[31] Anna Clara Monti. "A proposal for a residual autocorrelation test in linear models". In: *Biometrika* 81 (1994), pp. 776–780.

[32] Tom Duchemin and Mounia N. Hocine. "Modeling sickness absence data: A scoping review". In: *PLOS ONE* 15.9 (Sept. 2020), pp. 1–10. DOI: 10.1371/journal.pone.0238981.

[33] C. R. L. Boot et al. "Prediction of long-term and frequent sickness absence using company data". In: *Occupational Medicine* 67.3 (Feb. 2017), pp. 176–181. ISSN: 0962-7480. DOI: 10.1093/occmed/kqx014. eprint: https://academic.oup.com/occmed/article-pdf/67/3/176/17696320/kqx014.pdf. URL: https://doi.org/10.1093/occmed/kqx014.

[34] Zahid B Asghar et al. "Trends, variations and prediction of staff sickness absence rates among NHS ambulance services in England: a time series study". In: *BMJ Open* 11.9 (2021). ISSN: 2044-6055. DOI: 10.1136/bmjopen-2021-053885. eprint: https://bmjopen.bmj.com/content/11/9/e053885.full.pdf. URL: https://bmjopen.bmj.com/content/11/9/e053885.

[35] A. Mavragani and K. Gkillas. "COVID-19 predictability in the United States using Google Trends time series". In: *Scientific Reports* 10 (1 Nov. 2020). URL: https://doi.org/10.1038/s41598-020-77275-9.

[36] Pauli Virtanen et al. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”. In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[37] Lars Buitinck et al. “API design for machine learning software: experiences from the scikit-learn project”. In: *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*. 2013, pp. 108–122.

[38] *Yahoo Finance*. Accessed on Feb 15, 2023. URL: https://finance.yahoo.com.

[39] “Hyperparameter selection”. In: (2021). URL: https://neuralprophet.com/guides/hyperparameter-selection.html.

[40] *Ökade möjligheter att jobba på distans*. Accessed on May 11, 2023. URL: https://www.trr.se/aktuellt/okade-mojligheter-att-jobba-pa-distans/.

[41] Public Health Agency Of Sweden. *Smittspridningen av covid-19, influensa och RS-virus ökar kraftigt*. Accessed on May 7, 2023. URL: https://www.folkhalsomyndigheten.se/nyheter-och-press/nyhetsarkiv/2022/december/smittspridningen-av-covid-19-influensa-och-rs-virus-okar-kraftigt/.

[42] Anne Miesperä, Sanna-Mari Ahonen, and Jarmo Reponen. “Ethical aspects of eHealth - systematic review of open access articles”. In: *Finnish Journal of eHealth and eWelfare* 5.4 (Dec. 2013), pp. 165–171. URL: https://journal.fi/finjehew/article/view/9401.