Nearest Neighbour-Based Data Augmentation for Time Series Forecasting

Duncan Wallace[1,2,3], Eoin Delaney[1,2,3], Mark T. Keane[1,2,3], and Derek Greene[1,2,3]

[1] Insight Centre for Data Analytics, University College Dublin, Ireland
[2] VistaMilk SFI Research Centre, University College Dublin, Ireland
[3] School of Computer Science, University College Dublin, Ireland

Abstract. Accurate forecasting in many industries, such as dairy production, is a key challenge to improve investment decisions, market planning, energy use, and environmental protection. As most real-world phenomena take place over time, many prediction tasks involve the use of time series or collections thereof. In this study, we propose a nearest-neighbour data augmentation approach which draws on information from a set of related time series. Given an input series, we identify other series which have similar profiles using a k-nearest neighbour approach, and use their data to augment the input series. This has the effect of both smoothing outlying observations and filling missing observations. We show that, by combining this method with a criterion to refine the cases for which augmentation is to be applied, overall forecasting accuracy can be improved. Our approach is evaluated on diverse real-world time series data from a number of different domains.

1 Introduction

Time series data has attracted a significant research effort in recent years, perhaps because of its centrality in accurate forecasting in a variety industry settings, from financial services to smart agriculture. Many different approaches to time series prediction have been proposed, along with different methods for data pre-processing and augmentation [1]. However, the reliability of real-world time series forecasting is often impacted by the presence of noise, incomplete data, and outliers [2]. In our work on developing precision-agriculture applications – on dairy production forecasting – we have encountered all of these problems. For instance, dairy production forecasting is based upon data that is influenced by external factors such as weather, calving rates and disease. This data is also subject to missing values arising from discrepancies in milk-collection practices and technical implementation difficulties [3]. Thus, methods to address these issues and thereby improve forecasting robustness constitute important undertakings.

One technique to enhance the performance of time series forecasting is the application of a k-nearest neighbours (k-NN) approach, which is typically applied to reduce noise and the impact of outliers on training sets. The use of k-NN can additionally improve training set quality when the volume of observations

is limited. Recent research has, to the best of our knowledge, used perturbed examples of existing observations to generate neighbours for the application of this approach, as opposed to using real-world occurrences [4][5][6]. However, in the context of time series analysis, for many applications we might have multiple distinct series, each corresponding to a different entity. For instance, in the context of milk production, we might have a large collection of series, each of which represents data for a different farm associated with the same dairy processor.

To harness this additional information provided for a collection of entities, we propose using an approach motivated by k-NN regression, which attempts to make a prediction for a continuous target by averaging cases in the same neighbourhood. In our case, we generate a forecast for a given input series by incorporating information coming from an aggregated set of forecasts generated for neighbouring entities. Specifically, we apply this approach using the popular Prophet forecasting algorithm [7] to explore its potential effectiveness in data smoothing, and to a lesser extent data interpolation. We perform a detailed evaluation using five real-world datasets from four different application domains. Our analysis considers the importance of both the neighbourhood size k and the volume of available data. Our findings presented in Section 4 indicate that not all data is suited to this augmentation approach. Therefore, we also provide a means to profile entities, using a validation window, to ensure that augmentation is applied to instances more suited its application. Using a combination of k-NN regression and this filtering criterion, we observe improved accuracy across all datasets, when compared to baseline forecasts produced by Prophet alone.

2 Related Work

The standard definition of time series forecasting assumes the existence of a set of time-ordered observations of a variable Y, denoted $y1, y2, \ldots yt$, where yi is the value of Y measured at time i, and defines the predictive task as trying to forecast the future values of this variable for time stamps $s > t$. Many variants of this general task exist, including the use of other measured variables as potential predictors of the future values of the target series Y. Still, the general assumption is that there is an unknown function that "maps" past observations to future values of Y. The learning goal is to approximate a function using some prediction error criterion and a historical record of observed values. The predictors used for forecasting the future values of Y are usually the most recent observations of Y, as the basic assumption of time series forecasting is that of the existence of some form of correlation between successive observations of the series. This is the methodology employed in most approaches to time series forecasting, including the well-known ARIMA models [8]. This strategy assumes that future values of the series are only dependent on a limited number of previous values.

High-dimensional multivariate data can sometimes produce superior forecasting results to those achieved using linear time series. For instance, in the context of single-farm prediction of milk production, the most successful models are the surface fitting model and the NARX (Nonlinear autoregressive model

with exogenous inputs (RMSE 75.5kg, 365 day horizon)) [9], using features such as Days-In-Milk (i.e., how many days a cow has been lactating) and the NCM (number of cows milked) in the herd. In this same farming context some studies have used as many as twelve features including genetics, feed and grazing management information of the individual cows. However, such high-resolution data is not readily available for most commercial farms. Consequently, multivariate time series forecasting will typically require domain specific approaches, and may exhibit scalability issues.

One popular time series forecasting procedure is Prophet [7], which is based on an additive regression model that includes a seasonal component and a piecewise linear or logistic growth curve trend. The platform includes seasonality effects for non-linear data and automatically detects changes in trends by selecting change points from the data. To this end it has had proven applicability in relation to large scale forecasting requirements and has outperformed ARIMA in a range of different tasks, including predicting COVID-19 cases numbers [10] and making long-range predictions in the context of animal disease spread [11].

The k-nearest neighbour algorithm is a well-known non-parametric method used for classification and regression. It has been labelled as a lazy learner as the algorithm does not learn a discriminative function, but rather stores the instances for later use. Given some features or explanatory variables of a new instance to be regressed on, k-NN finds the k training instances that are closest to the new instance according to some distance metric.In the context of time series data, k-NN has been employed in a range of disparate ways. For instance, its use can be witnessed in local prediction: by breaking down domains into local neighbourhoods and fitting to each neighbourhood separately. This type of methodology can be found in bootstrapping as far back as Jayawardena and Lai [5], or more recent variants such as that found in Wu at al.'s paper concerning support vector regression, where the weighted averaging of k neighbours is adopted and the weight function made proportional to the inverse of square Euclidean distances between $Y(t)$ and $Y(t0)$ [4]. In particular, many papers specifically consider time series nearest neighbours from a position of a lack of extant real-world neighbours within their datasets. As noted by Martínez et al., there has been limited application of k-NN regression to time series forecasting [6]. This may be specifically exacerbated by a dearth of datasets in some domains that provide collections of related time series for k-NN to be performed upon. However, later in Section 3.2, we highlight a number of domains where collections of related series are readily available.

3 Methods and Materials

3.1 Proposed Method

A variety of techniques have been proposed for manipulating time series representations to improve the outputs of downstream tasks. While *data augmentation* has largely been used to generate additional training data in the context of image datasets, Oh et al. [12] considered producing additional time series for

classification, using an interpolation-based approach. In the context of time series forecasting, a more common approach for producing robust estimates has been to construct *ensembles* of multiple models [13, 14]. Such approaches can either be used to improve the performance of a group of models, or reduce the likelihood of an unfortunate selection of a poor model. A feature of many real-world data sources is the presence of both noise and missing values. Therefore, we employ two different augmentation methodologies to improve predictive performance in such an environment. We use the k-NN algorithm in the form of data augmentation of data series to reduce noise and erratic values, while k-NN for data interpolation is also performed to fill in missing values for a given series.

Data augmentation. Given a collection of related time series $\{T_1, \ldots, T_n\}$, representing n different entities, we produce a forecast for an individual series T_i as follows. Firstly, we identify its k nearest neighbouring entities using unweighted Euclidean distance over time series of equal length. Next, we produce an augmented time series for each entity that comprises the arithmetic mean of observances, or interpolated instance in the event of a missing value. The arithmetic mean of time series was chosen as augmentation method as the fluctuations between the real world datasets chosen (such as meteorological or dairy production) were on a linear rather than logarithmic scale. The resulting augmented series $T_i{}'$ was then used as the input data for a standard forecasting model. In our case, we employ Prophet, which has, to be best of our knowledge, never been tested with k-NN augmentation.

Selection criteria. Early experimentation indicated that data augmentation was not universally appropriate for every entity in a dataset. Consequently, we have developed an approach for filtering out entities that are less suited to augmentation. These might correspond to entities for which forecasting accuracy is already high, or outliers for which relevant neighbouring cases are unavailable. We can consider the time-series to be composed of distinct windows, with the term 'window' referring to a slice of the time series (which in the given datasets corresponds with a single year). Using a single window of the data as a validation set, we apply augmentation for a range of values of k, followed by the forecasting model in each case. We can then determine whether applying k-NN resulted in any positive change in forecasting performance (e.g., as measured by mean absolute error (MAE)), and, if it did, what value of k produced the best result. If no value of k produced a positive change then the entity is filtered from the augmentation process. We refer to the entities for which augmentation is applied as "contested", and those for which it is not applied as "uncontested".[4] However, if the approach produced a positive result when using only one value of k, an

[4] Given that we are interested in the effectiveness of both the filtering technique, and data augmentation, we consequently adopt this terminology to add clarity and distinguish between the original results that form our baselines, which are entirely unaugmented, and the test results which consist of both augmented and unaugmented cases.

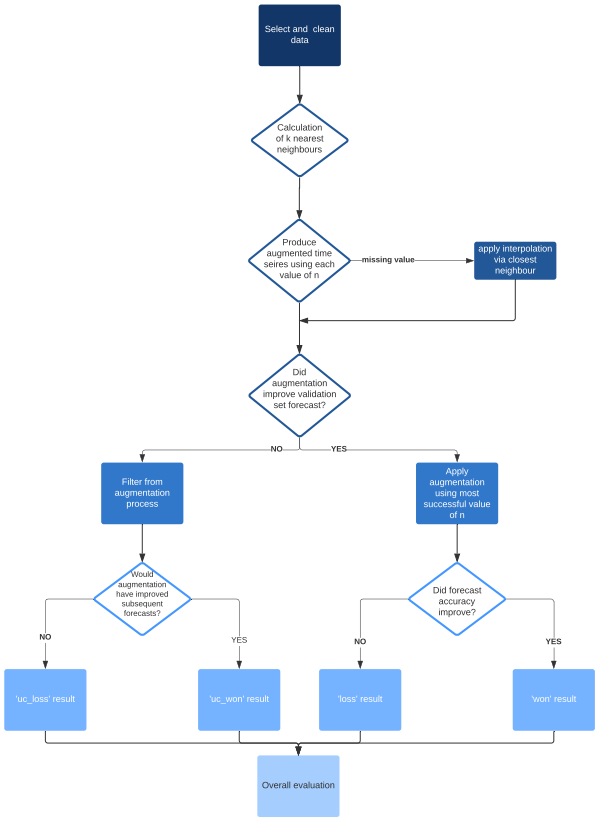

Fig. 1. Overview of methodology and evaluation steps employed.

additional check was performed to see if the difference in MAE for this entity's validation window was above average (across all entities for the validation window). If this happened to be the case, then such entities were also discarded for use with the proposed approach. A comparison between entities which were and were not selected using this basis will be provided in Section 4.

3.2 Datasets

We now discuss five different datasets from domains where collections of related entities, represented as time series, naturally occur. We chose a diverse range of domains in order to prove the applicability of our approach. With the exception of the socioeconomic dataset, each of these domains is represented as univariate time series (while the lattermost, for the purposes of this research, is split into two separate univariate datasets).

Agricultural data. Our first dataset concerns milk production from 400 farms in Ireland for one dairy processor, as measured across four consecutive years. This

will be referred to as the 'dairy' dataset in our experiments. Cases describe daily readings for the volume of milk collected on each farm. Each farm constitutes a separate entity and the predicted variable is the volume of milk collected from each entity. In any given year, milk supply forecasting is a fundamental driver for the dairy sector. Dairy processors use their forecasts to establish pricing levels, contracts with supplier farms, and the production requirements for their factories. Forecasting consequently strongly influences farm management (the prevention of under or over production, and also the manner of production adopted), the consumption of resources in the sector (e.g. fertiliser use, tanker-transport use for milk collections), and the processing efficiency of factories (avoiding waste from surplus milk supplies). Processors can drive sustainability changes through accurate and precise forecasting. However, such forecasting is highly challenging for many reasons: the forecast must be made over 1000s of farms which differ in their herd-profiles; the land farmed and farm-management practices is for the full year in advance, not incrementally as the year unfolds; and planning can encounter disruption from climate-change and disease outbreaks.

Meteorological data. The second dataset used here, involves the monthly average temperatures, recorded in degrees Fahrenheit, of 18 major cities from regions spread across the world, dating from the years 2000 and 2013. This corresponds to a subset of the Berkeley Earth climate change dataset [15]. This publicly available data, referred to as the 'weather' dataset in our experiments, concerns locations ranging from a latitude of 59.33 North to 37.78 South [16]. With values of southern latitudes considered negative, the mean latitude was 24.06 with a standard deviation of 39.3. Entities in this dataset represent individual cities, and the forecast variable is the monthly mean temperature of the given entity, over a window of a particular year. This type of mid- to long-range meteorological forecasting based upon temperature is very important for many economic sectors While the time-scale used in the course of this research was too small for longer climate-trend observation, improved methods to towards this end also have a bearing on the analysis of wider climate change modelling.

Socioeconomic data. The third dataset relates to the Local Area Unemployment Statistics (LAUS) program of the Bureau of Labor Statistics (BLS) of the United States. This consists of monthly estimates of civilian non-institutional population, labour force participation rates, and employment-population ratios for the 50 states and the District of Columbia from January 1976 onwards. Two separate datasets for this exist: one which is seasonally-adjusted (excluding civilian non-institutional population ages 16 and older) and one which is not (raw number of unemployed). These two datasets will be referred to as 'employment adj' and 'employment' respectively within tables. The seasonally-adjusted figures are determined by the BLS using the seasonal components of the LAUS labour force. For our purposes we used a subset of both these data, representing a range of 13 years, from which we obtained both our training and testing sets [17]. These two datasets consider states to be individual entities with the fore-

Table 1. Summary of datasets used in our experimental evaluations.

Dataset	Total Entities	Test Windows
dairy	400	2
weather	18	7
employment	40	10
adj employment	40	10
gas	18	7

cast variable being the number of unemployed people within each state (which, depending on the dataset, may have been seasonally adjusted). Forecasting of unemployment figures (particularly non-seasonal) is difficult due to the large number of external factors that have a bearing on these numbers. Sudden increases in non-seasonal unemployment often represent economic shocks triggered by unpredictable macroeconomic events. Being able to accurately forecast unemployment numbers therefore has a bearing on the capacity to predict recessions and larger economic cycles.

Energy data. The fifth and final dataset considered relates to the Natural Gas Gross Withdrawals and Production, measured in monthly million cubic feet per day, as compiled by the U.S. Energy Information Administration (EIA) [18]. This encompasses 18 separate entities (16 US states, the Federal Offshore production in the gulf of Mexico, and a conglomeration of all other US territories). This dataset will be referred to as 'gas' in our experimental results. The forecast variable relates to the gross withdrawal amount. Despite being subject to arguably fewer external factors than some of the earlier discussed datasets, accurate forecasting of gas production is nevertheless very problematic. Gas production is not entirely insulated from economic processes, and is dependent on a resource the volume of which is often difficult to estimate and whose extraction is non-trivial. Political policy may also have a significant bearing on production levels (particularly in recent years on the subject of fracking). All of these aspects contribute to significant issues when attempting to provide accurate forecasts.

4 Experimental Evaluations

4.1 Method

Measures. Using the datasets described earlier we evaluated our proposed method using several measures. The principle measure used was the mean absolute error (MAE), in which a lower figure indicates a better performance. The baseline method in the experiments is the non-augmented Prophet forecast model applied to a single query time series.

Setup. For the purposes of our experiments, the first three windows in each dataset were used as training data, the next window as a validation set, and the

subsequent windows as test data (the latter of which is used for Tables 3-4). For each window of the test data, we have a predicted and actual value. Forecasts were generated for each entity to provide a base rate and then further forecasts generated using training data composed as as an arithmetic mean for a range of neighbours. The base rate in these evaluations was the non-augmented predicted value provided by Prophet. The window of observations and predictions was then advanced by one period, with the length of a period being equal to a calendar year. The process was then repeated again *ab initio*, until the methodology was run over the full length of the respective dataset.

A potential limitation when using Euclidean distance as a metric to measure time series similarity is that when large absolute distances exist between different time series, it becomes harder to identify time series with similar trends and seasonality. In order to deal with this issue, we applied a normalisation in which the time series being compared were scaled, in proportion to the ratio of means. Missing data can also be an issue for forecasting in real-world datasets. While a potential solution to this can be to fill a missing value with the preceding day's value, in our method we instead replace the missing value with a mean of the values of k-nearest entities (i.e., interpolation).

Prophet has no native means to handle multiple entities' time-series-data for fitting an individual entity's time series forecast, so our proposed approach is consequently applied prior to the provision of data to the model. Prophet allows the use of a custom list of holidays and seasonalities in the model. However, the use of holidays as a parameter is only appropriate for time series whose values are discredited to days as opposed to months or years. As only one of our datasets contained daily data (the agricultural dataset), and the importance of holidays in this context was unclear, we do not utilise this functionality in our forecasts.

4.2 Results and Discussion

Number of neighbours. In our experiments, we consider augmentation using $k \in [1, 5]$ neighbours. As outlined in Section 3.1, the validation window was used to establish whether or not the baseline forecasts would be "contested" using the data augmented via the proposed k-NN approach (i.e., whether or not additional neighbours should be used). However, the validation window is also used to establish what value of k should be used to this end. A frequency count of the value of k which produced the best result within the validation window was performed for all entities which passed the selection criterion. The results of this for each dataset can be viewed in Table 2. When the number of entities to provide this summation of k values was at least double the range of k values available (i.e., a clear majority), a global value of k was used for all further forecasts for that dataset (otherwise a local value of k was used).

Table 3 shows the difference in performance between these forecasts when the uncontested baseline forecasts are used, versus when the selection criteria determines that the augmentation approach should be used. In the table, *uc_lost* and *uc_won* refer to the uncontested cases which have not been augmented, and whether or not the proposed approach would have produced superior or inferior

Table 2. Optimal k value for selected entities on each dataset.

	k				
Dataset	1	2	3	4	5
weather	0	1	2	2	2
dairy	11	19	33	59	**111**
employment	**9**	3	2	4	6
employment (adjusted)	**11**	3	2	4	7
gas	1	2	0	1	1

Table 3. Summary of wins and losses for all entities in each dataset where the proposed method is applied (won and lost) and where it is not applied (uc_won and uc_lost).

Dataset	uc_won	uc_lost	won	lost
weather	36	107	35	20
dairy production	106	74	138	82
employment	72	90	87	84
employment adj.	76	104	87	93
gas production	11	79	11	7

Table 4. Summary of mean absolute error (MAE) performance on each dataset.

Dataset	uc_won	uc_lost	won	lost
weather	43.43	-176.88	22.32	-6.69
dairy production	26843.21	-18208.28	36671.28	-19619.10
employment	275783.67	-2403315.51	764489.92	-687292.87
employment adj.	505932.30	-2987825.88	764489.92	-737828.20
gas production	1194.00	-93314.10	3292.11	-2866.42

results had it been used. These results relate to the test data of all chosen entities, where the maximum value is the number of windows multiplied by the number of entities for a given dataset. Where the augmentation approach is adopted we can also see, for a given forecast, whether or not the change has resulted in an increase or decrease in accuracy (i.e., won and lost respectively).

Table 3 reveals considerable differences between datasets. For instance, the dairy dataset would have seen overall more forecasts improve than deteriorate through k-NN without augmentation applied to its data, but the weather dataset would have seen two thirds of its non-augmented forecasts obtain worse results than that of the baseline. Overall, the proposed selection criteria appears to have achieved a reliable basis for excluding many of the forecasts less suited to an augmentation approach. Indeed, only one dataset (employment adj.) shows more cases chosen by the selection criteria suffering from reduced accuracy than the number of those benefiting.

Moving beyond a simple count of wins and losses to the overall magnitude of these changes (Table 4), again we observe that uncontested forecasts, which fail

Table 5. Summary of overall results, using the proposed selection criterion.

	weather	dairy	employment	employ. adj.	gas
Mode for k	-	5	1	1	2
Records	55	220	171	180	18
Mean MAE	3.84	787.18	26453.74	25919.66	528.64
Mean MAE change	-0.28	-77.51	-451.44	-148.12	-23.65

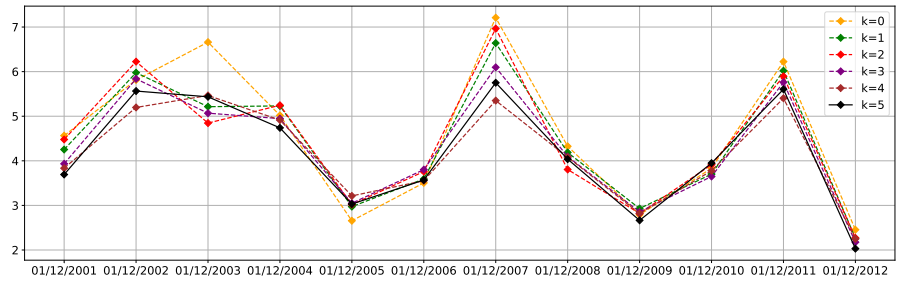

Fig. 2. Temperature forecast accuracy over holdout windows for the entity Warsaw, for values of $k \in [1, 5]$, (relating to each value's respective MAE value), with the baseline indicated by '$k = 0$', and MAE represented by the y axis.

the selection criteria, in general would have seen a performance decrease with the application of k-NN. Overall, we can see that there is not a single dataset in which the overall MAE declines for the selected forecasts. An overall summary of this information is given in Table 5. We report both the average MAE and the average reduction in error across entities that conform to the selection criteria.

If we look at a single sample entity in Fig. 2 that qualifies for the application of the augmentation approach in the weather dataset (specifically depicting Warsaw, with y-axis representing MAE value), we can see the approach's accuracy relative to both the baseline Prophet MAE and the hypothetical performance using different values of k. Using the validation window corresponding to 2001, all values of k showed improvement over the baseline Prophet forecast, and as the weather dataset applies a local rather than universal value for k on the test set, the value of k that performed best in this instance was chosen ($k = 5$). While the forecasting performance for all values of k broadly match the peaks and troughs of the baseline, significant improvement can be seen across the different windows. In this case the nearest neighbours which combine to provide the augmentation training data for MAE5 are, in ascending order: Wroclaw, Kiev, Kherson, Stockholm, and Uppsala.

Fig. 3 shows augmentation performance for a single farm from the dairy dataset where x axis represents time, and y axis represents volume of milk produced in litres. In the plot we can see how close the different augmentation approaches ($\hat{y}_1$-$\hat{y}_5$), as well as the baseline Prophet forecasting ($\hat{y}$) come to predicting the actual production values (y) for this window. As the dairy dataset

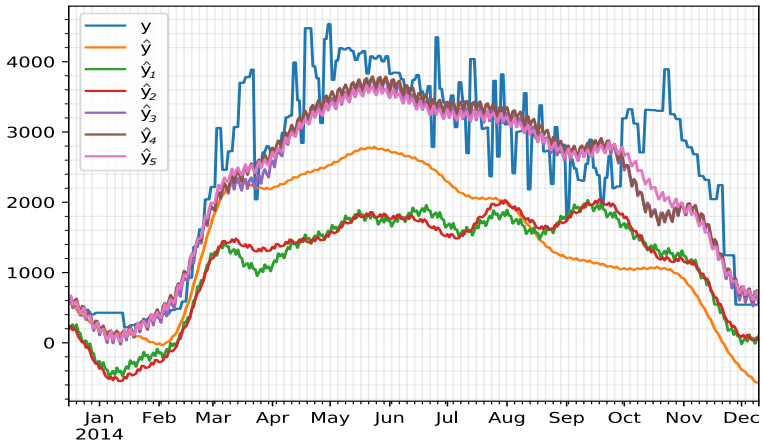

Fig. 3. Monthly milk production forecasts generated using different k-NN augmentation approaches for values of $k \in [1, 5]$, with the baseline indicated by $\hat{y}$ and ground truth indicated by y.

utilised a global value of k, $\hat{y}_5$ was the model chosen for all entities that were selected for augmentation. Notably, not all values of k in this instance achieve improved results over the baseline, with both $\hat{y}_1$ and $\hat{y}_2$ being generally inferior to the baseline. However, $\hat{y}_5$ provides a good match for the actual production values of this farm.

5 Conclusions and Future Work

Despite the strong performance of the baseline Prophet algorithm, our findings indicate that potential improvements can be made with an aggregation approach to the training data using a k-NN methodology, in situations where we have multiple related entities represented as time series. Our results indicate that, while this has demonstrable benefit, it is by no means universally successful. Nevertheless, there is clear promise offered by using different time series with similar profiles, in domains where forecasting is challenging due to noisy or missing data.

Although this work has established the potential use of this methodology, questions remain concerning the potential use of different distance metrics for the measurement of neighbours, and whether the use of additional validation windows could provide sufficient data to adapt a more precise aggregation technique. While the various datasets explored provided a diverse ranges of entities, it was beyond the scope of this research to establish how many entities are required for this methodology to be effective. Furthermore, we intend to extend this research by considering how k-NN in this context can ultimately aid in providing use explanations of forecasting outputs to end users.

Acknowledgements. This publication has emanated from research conducted with the financial support of (i) SFI and the Department of Agriculture, Food

and Marine on behalf of the Government of Ireland to the VistaMilk Research Centre under Grant Number 16/RC/3835, and (ii) SFI to the Insight Centre for Data Analytics under Grant Number 12/RC/2289_P2.

References

1. Ganapathy Mahalakshmi, S Sridevi, and Shyamsundar Rajaram. A survey on forecasting of time series data. In *Proc. ICCTIDE'16*, pages 1–8. IEEE, 2016.
2. Chris Chatfield. *Time-series forecasting*. CRC Press, 2000.
3. Nicolas Urruty, Delphine Tailliez-Lefebvre, and Christian Huyghe. Stability, robustness, vulnerability and resilience of agricultural systems. a review. *Agronomy for sustainable development*, 36(1):15, 2016.
4. K.W. Lau and Q.H. Wu. Local prediction of non-linear time series using support vector regression. *Pattern Recognition*, 41(5):1539–1547, 2008.
5. AW Jayawardena and Feizhou Lai. Analysis and prediction of chaos in rainfall and stream flow time series. *Journal of hydrology*, 153(1-4):23–52, 1994.
6. Francisco Martínez, María Pilar Frías, María Dolores Pérez, and Antonio Jesús Rivera. A methodology for applying k-nearest neighbor to time series forecasting. *Artificial Intelligence Review*, 52(3), 2019.
7. Sean J Taylor and Benjamin Letham. Forecasting at scale. *The American Statistician*, 72(1):37–45, 2018.
8. Ana Maria Bianco, M Garcia Ben, EJ Martinez, and Vıctor J Yohai. Outlier detection in regression models with ARIMA errors using robust estimates. *Journal of Forecasting*, 20(8):565–579, 2001.
9. Fan Zhang, Michael D Murphy, Laurence Shalloo, Elodie Ruelle, and John Upton. An automatic model configuration and optimization system for milk production forecasting. *Computers and electronics in agriculture*, 128:100–111, 2016.
10. Arko Barman. Time series analysis and forecasting of covid-19 cases using LSTM and ARIMA models. *arXiv preprint arXiv:2006.13852*, 2020.
11. Cong Xie, Haoyu Wen, Wenwen Yang, Jing Cai, Peng Zhang, Ran Wu, Mingyan Li, and Shuqiong Huang. Trend analysis and forecast of daily reported incidence of hand, foot and mouth disease in hubei, china by prophet model. *Scientific reports*, 11(1):1–8, 2021.
12. Cheolhwan Oh, Seungmin Han, and Jongpil Jeong. Time-series data augmentation based on interpolation. *Procedia Computer Science*, 175:64–71, 2020.
13. Mariana Oliveira and Luis Torgo. Ensembles for time series forecasting. In *Asian Conference on Machine Learning*, pages 360–370. PMLR, 2015.
14. Shaolong Sun, Yunjie Wei, and Shouyang Wang. Adaboost-lstm ensemble learning for financial time series forecasting. In *International Conference on Computational Science*, pages 590–597. Springer, 2018.
15. R. A. Rohde and Z. Hausfather. The berkeley earth land/ocean temperature record. *Earth System Science Data*, 12(4):3469–3479, 2020.
16. Climate change: Earth surface temperature data. `https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data`. Accessed: 2021-01-30.
17. States and selected areas: Employment status of the civilian noninstitutional population, january 1976 to date. `https://www.bls.gov/web/laus/ststdsadata.txt`. Accessed: 2021-01-30.
18. Natural gas gross withdrawals and production. `https://www.eia.gov/dnav/ng/hist/ngm_epg0_fgw_nus_mmcfdm.htm`. Accessed: 2021-01-30.