STAR ENSEMBLE: A NOVEL ALGORITHM FOR SPATIO-TEMPORAL DATA DECOMPOSITION AND INTERPOLATION.

James Monks and Liwan Liyanage

School of Computing Engineering and Mathematics, Western Sydney University, Australia

ABSTRACT

Spatio-temporal data is becoming increasingly prevalent in our society. This has largely been spurred on from the capability of building arrays and sensors into everyday items, along with highly specialised measuring equipment becoming cheaper. The result of this prevalence can be seen in the wealth of data of this kind that is now available for analysis. This spatio-temporal data is particularly useful for contextualising events in other data sets by providing background information for a point in space and time. Problems arise however, when the contextualising data and the data set of interest do not align in space and time in the exact way needed. This problem is becoming more common due to the precise data recorded from GPS systems not overlapping with points of interest and not being easily generalised to a region. This is Interpolating data for the points of interest in space and time is important and a number of methods have been proposed with varying levels of success. These methods are all lacking in usability and the models are limited by strict assumptions and constraints. This paper proposes a new method for the interpolation of points in the patio-temporal scope, based on a set of known points. It utilises an ensemble of models to take into account the nuanced directional effects in both space and time. This ensemble of models allows it to be more robust to missing values in the data which are common in spatio-temporal data sets due to variation in conditions across space and time. The method is inherently flexible, as it can be implemented without any further customisation whilst allowing for the user to input and customise their own underlying model based on domain knowledge. It addresses the usability issues of other methods, accounts for directional effects and allows for full control over the interpolation process.

1. INTRODUCTION

Data on discrete spatio-temporal events can be enriched through the inclusion of ancillary data occurring at the exact point in space and time. Unfortunately, useful background data is often not measured at the exact point that is needed to inform the event. In order to generate more relevant data, point based interpolation can be performed, using the observed points as inputs to predict for the needed points. A motivating example for this can be found in the relationship between air quality and asthma attacks. Records of emergency cases of asthma are recorded based on ambulance response, or self reported location/time of attack. This data could then be utilised along with pollution and weather data to understand the impact of air quality and conditions on the rate/severity of asthma attacks. The issue with this is that monitoring stations that are measuring these conditions, will not be located at the site of asthma attacks, nor are the attacks likely to fall exactly on schedule with the record time intervals.

Natarajan Meghanathan et al. (Eds) : ACITY, AIAA, DPPR, CNDC, WIMNET, WEST, ICSS - 2019
pp. 73-79, 2019. © CS & IT-CSCP 2019 DOI: 10.5121/csit.2019.91707

1.1. Related Work

A literature search has been carried out into spatio-temporal interpolation that informs the methodology and implementation. These methods are derived from spatial interpolation and are extended to incorporate time as an additional dimension (Eldrandaly 2017). This is done through reduction techniques which treat the problem as a series of spatial interpolations, or extension techniques, which treat the time component in the same way as the spatial ones. The issue of computational complexity present in both approaches and needs to be considered when proposing new methodology. Shape functions, IDW and Kriging methods (Cressie& Hawkins 1980) are commonly used (Pebesma & Heuvelink 2016). The methods proposed in this paper describe an alternate framing of these problems, through decomposition of the spatial and temporal components and an ensemble of traditional models on these componetns.

2. METHOD

2.1. Key Concepts

Observations that are close together are more similar than observations that are a great distance apart. This is true for separation distances across both space and time. This result allows an autoregression to be constructed on the number of lags on this separation. It is trivial to set up an autoregression on time lags, however, the same cannot be said for spatial lags as a simple distance metric would lose important directional information. This loss of directional information extends to separation distanced of both space and time (i.e. a different location at a different time). This means that any relationship that is not independent spatially and spatio-temporally, will not be accurately modelled with this simple application of autoregression.

2.2. Star Ensemble 2D

The directional effects will first be addressed in 2 dimensions, which can be thought of as latitude and longitude for the sake of application, but will be referred to as the x, y plane for explanation. Consider an unknown point **p** in this plane, with known valued point$\boldsymbol{p_m}$s spaced at irregular intervals in both x and y directions. This data is then filtered to contain only those points within some given radius from $\boldsymbol{p}$ due to relevance (based on knowledge of the data). The plane is then divided evenly by $\boldsymbol{n}$ lines that pass through $\boldsymbol{p}$. $\boldsymbol{N}$ divisions of the data are created by filtering the known points $\boldsymbol{p_m}$for each line, based on the perpendicular distance between the point and the line being less than some number $\boldsymbol{k}$. The closest point point on the line to each $\boldsymbol{p_m}$is denoted as $\boldsymbol{q_m}$. A weighted regression is then set up for each of $\boldsymbol{n}$ divisions to model the values given at known points $\boldsymbol{p_m}$and the distance between $\boldsymbol{q_m}$and $\boldsymbol{p}$, with weights defined as the distance between $\boldsymbol{p_m}$and $\boldsymbol{q_m}$. Each of these $\boldsymbol{n}$ models is used to predict for the value at point $\boldsymbol{p}$, with a weighted average of each of the predictions being used to generate a final prediction for this point. The weights for this weighted average are based on the RMSEs of each of the models.

2.3. Spatio-Temporal Interaction

The previous process is concerned with identifying spatial directional trends, however, these concepts can be extended to cover spatio-temporal directional trends. This is performed through first applying the process using planes angled at regular intervals (instead of lines) to divide a sphere. After the divided data sets are created, each of the points on the plane can be considered in the same way as in the original process, resulting in a number of directional trends that are along the spatio-temporal directional components identified. This is formalised below.

2.4. Spatio-Temporal Star Ensemble.

The steps in 2.3 are repeated for the xy plane, without models being created. This utilises points in 3 dimensions xyz (with z representing time in this instance), and maps the points pm onto the closest point on each line $\boldsymbol{q_m}$. The resultant evenly spaced lines through xy are representative of planes through xyz. For each of these planes, the process of dividing data by drawing lines through a centre point (along the plane) is used, to establish $\boldsymbol{n}$ lines along each of m planes. This results in $\boldsymbol{m \cdot n}$ lines through the 3 dimensional space xyz, which can be considered directional components space and time. The divisions of data are then created based on whether the distance between each of the known valued points pm and the closest point on the line $\boldsymbol{q_m}$ is less than a specified threshold. This is repeated for every line.

Once the divisions of data have been created, a model is constructed from each line through xyz. This is done in a similar way to the 2d version of this process in that the value at a point is predicted for based on the distance between the closest point on the line of each point $\boldsymbol{q_m}$and the point $\boldsymbol{p}$. The distance between each pm and $\boldsymbol{q_m}$is then used to inversely weight the model that is being created, as the further away these points are from the trend of direction, the less influence they will have on the model. After these models have been created, a weighted average of their predictions based on RMSE is used to evaluate the final result.

3. APPLICATION

An R package has been developed as an implementation of this process. This is can be easily applied to any given point in space and time, in the same way as is common to predict values in R. There is no training data involved as the framework uses all available relevant data to predict for a central value. All results that are used following have been calculated using this package.

3.1. Climate Data Integration

This process of predicting inward based on 3 dimensional is particularly important in the integration of spatio-temporal data from different sources. This process allows for data measured at different, irregular locations at different time intervals to be integrated into the one data set by predicting realistic values of the contextual data at the point of interest. This integration allows for important contextual information to be analysed alongside events, allowing analyses on pollution and health, weather and crop yield, etc to be conducted with relative ease. A simple interface in the form of an R package is provided for easily integrating contextual data to the reference events.

The data on atmospheric conditions is extremely important in understanding the impact of environmental effects on asthma attacks and other respiratory events. A traditional integration of atmospheric data and asthma data would consist of identifying the closest atmospheric observation to the point at which the asthma attack occurred and assigning this observation to the attack. This is prone to error when there is a large separation distance or when the atmospheric observations are irregularly spaced. The method being put forward would instead identify directional trends in the surrounding contextual data and create a prediction based on an ensemble of models for the point in space and time of the asthma attack.

This model of identifying an event and applying the proposed process to predict the contextual data for the event is the ideal application of the method. The simple interface allows for data to be provided to the model as reference events (in this case asthma). The contextual data is then provided in whichever structure it is in, with latitude, longitude and time columns being essential. The process then integrates the contextual data with the event data of interest and

returns a single observation for each event. This returned data can be customised to include lags and leads of contextual data for time series analysis if required.

An important note about the data collected from both weather and pollution monitoring stations is that there is an inherent issue with missing data. Even data collected on the same type of condition across monitoring stations that are distributed spatially has this issue due to differing collection practices, observation periods and temporary equipment failure. An important aspect then becomes the sensitivity or lack thereof of the integration model.

4. EVALUATION

This method will be evaluated through using collected atmospheric conditions data. A set of data describing the pollution in Victoria at half hour intervals will be used, along with a data set describing the daily temperatures in NSW. This is real data that is to be used to inform future support systems in these areas, after it is integrated with health data using the methods described. The temperature and pollution data sets have been chosen, as they are representative of different phenomena that is likely to be modelled. In particular, temperature varies relatively smoothly spatially and temporally (i.e. it is rare for extreme drops or spikes in temperature), especially when it is aggregated daily by taking the maximum. In contrast pollutants have the potential to move in high concentration masses or clouds. This means that it would be common to see extreme changes happen quickly, which can be observed at a high granularity through the half hourly measurements.

The method is evaluated against methods that are commonly in use for this type of problem. Namely these methods are the nearest neighbour method and the K Nearest Neighbours approach. It is not being compared to spatio-temporal kriging or similar methods as these methods assume spatio-temporal stationarity and are not readily applicable to irregularly spaced data with missing values. This evaluation is performed through taking the existing weather and pollution data sets, removing a sample of the stations for all time points and filtering out all of the sampled points in which there is a missing value. This validation set is predicted for using the remaining data and the values are compared to generate a performance metric. In this case the mean absolute error will be used so as to not place undue weight on outliers.

5. RESULTS

5.1. Max Temperature

The starensemble model was able to predict large portions of the temperature validation set accurately. The performance is comparable to that of both the nearest neighbour interpolation and the K Nearest Neighbours interpolation (Table 1). The proposed method did not perform quite as well as the nearest neighbour class of methods, however, there were a few predictions that were very far off dragging this error metric up (figure 1).

TABLE 1. Model mean absolute errors for max temperature data

Method	MAE
Star Ensemble	4.200181
1 Nearest Neighbour	3.704762
5 Nearest Neighbours	3.22662

It can be seen in table 2 that the star ensemble method is robust to failed predictions resulting from missing values. The nearest neighbour method was particularly sensitive to this problem

and the K nearest neighbours method was slightly sensitive, though not nearly to the extent of the nearest neighbour prediction. This is important as it means the starensemble model maintains a high predictive accuracy while making these previously impossible predictions.

TABLE 2. Failed predictions for each model

Method	Failed Predictions	Attempted Predictions	Percent Failed
Star Ensemble	0	109	0%
1 Nearest Neighbour	25	109	22.9%
5 Nearest Neighbours	1	109	0.917%

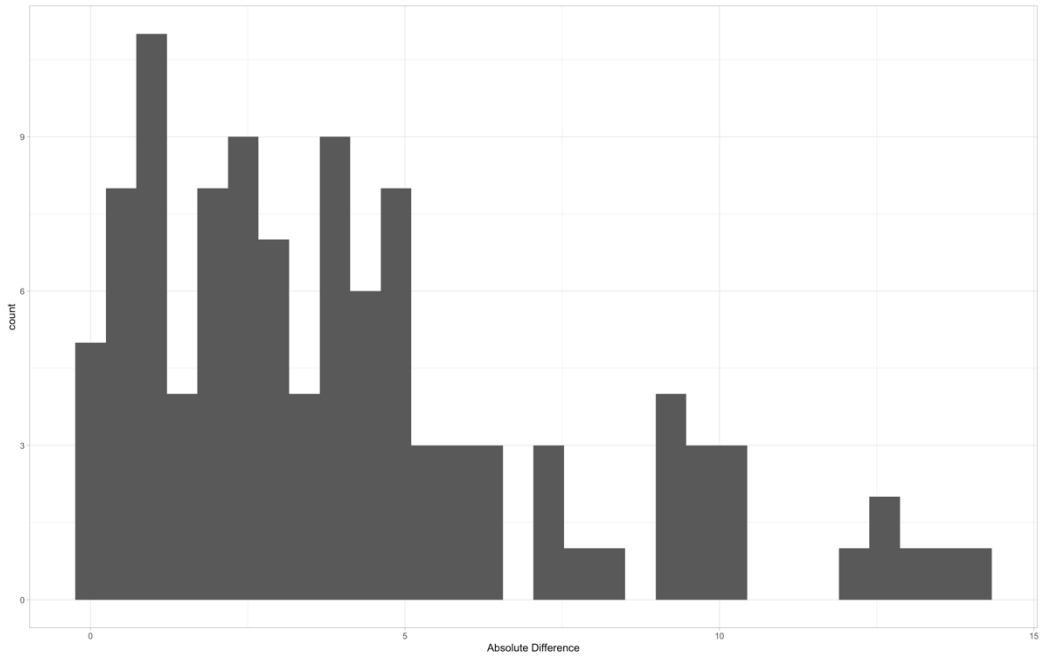

FIGURE 1: Histogram of absolute errors for the max temperature data

O3 Pollution

The predictive capacity of this method when applied to the pollution data is still comparable, though not nearly as good as it is in the case of the temperature. It is interesting to note that the error actually goes up when increasing the number of nearest stations used in the prediction in this case, though only slightly.

TABLE 3. Model mean absolute errors for O3 pollution data

Method	MAE
Star Ensemble	7.891094
1 Nearest Neighbour	3.327586
5 Nearest Neighbours	3.335135

The star ensemble method is again robust to failed predictions. The nearest neighbour class of predictions however are also less sensitive to this issue. This may indicate there are fewer

missing values in this data set than in the temperature one, or that they occur less commonly together.

TABLE 4. Failed predictions for each model

Method	Failed Predictions	Attempted Predictions	Percent Failed
Star Ensemble	0	185	0%
1 Nearest Neighbour	11	185	5.95%
5 Nearest Neighbours	0	185	0%

6. DISCUSSION

The proposed method shows a comparable performance to the methods that are commonly used. It performs better in the case of predicting the maximum temperature than in the case of the pollution concentration. This shows that there is merit to this style of spatio-temporal decomposition, however it needs to be improved in order to handle extreme cases better.

The worst predictions that are made by the model are often predicted exceedingly well by the nearest neighbour class of approaches. The opposite is also true that the poor predictions made by the nearest neighbour based approaches are predicted well by the proposed method. This indicates that there is a potential to create a composite algorithm that is even more accurate. It is also potentially indicative of a way to improve the decomposition modelling through accounting for these situations. This could take the form of placing more weight on the closest points for given situations.

The biggest advantage over the previously used models in this case is the performance on the observations that contained missing values. This is important as missing values such as these are inevitable when utilising data from a variety of sources distributed across space and time.

This algorithm proposes a unique and extensible way of analysing interesting spatio-temporal patterns. It does not solely consider the spatial effects or the temporal effects, rather it considers the variety of spatio-temporal trends and interactions. This method can be applied to any regular or irregular pattern of spatio-temporal data and has the potential to be extended through unique decomposition strategies.

6.1. Limitations and Future Work

This method is relatively resource intensive as the partitions need to be made for each data point. It is not so slow as to be a significant hindrance in on-time data integration operations, however it is not currently feasible to utilise this method in a production environment. It is possible to run these computations in parallel however, as there are no dependencies on other points in each prediction calculation. This would result in a significant speed boost depending on the number of processes that it is being run on.

The models built for each segment can fit values for anywhere within the segment, not only the centre point as is currently used in this application. It is possible to create a field of predictions through using overlapping segmentation areas. This is opposed to the point based predictions that are currently being implemented. If this is coupled with parallel processing, there is the potential for this to be used in production, as many of the calculations could be shared across predictions.

The assumption that is underlying the decomposition is that the phenomena being predicted is distributed in a smooth gradient. This is the case with conditions such as temperature, however, for conditions such as precipitation or pollution, it may be more accurate to assume that there exist objects with sharper boundaries. In these examples, these objects could be thought of ads rain clouds or masses of gas. Decomposing based on the detection of such an object presence or absence may allow for a better encapsulation of different levels of variation in a specific region.

The current implementation uses a locally weighted regression model to predict the value of a variable. This prediction accuracy of the overall method may be improved through utilising time series methods due to the temporal affect on the segmented data sets. Non-parametric methods could also be used such as decision trees or random forests. This would not be hard to implement due to the modularity of the proposed framework.

REFERENCES

[1] Pebesma, E & Heuvelink, G 2016, 'Spatio-temporal interpolation using gstat', RFID Journal, vol. 8, no. 1, pp. 204-18.

[2] Eldrandaly, K & Abdelmouty, A 2017, 'Spatio-temporal interpolation: Current Practices and Future Prospects', International Journal of Digital Content Technology and its Applications, vol. 11.

[3] Cressie, N & Hawkins, DM 1980, 'Robust estimation of the variogram: I', Journal of the International Association for Mathematical Geology, vol. 12, no. 2, pp. 115-25.