Dissertation Proposal:

Improving Spatio-Temporal Forecasting by Extracting Indicators Based on Past Observations

J. Rapson
University of Oxford Department of Statistics

February 1, 2023

1 Supervisors

Sir Bernard Silverman, FRS
Emeritus Professor of Statistics, University of Oxford
Chair, Geospatial Commission
Former Chief Scientific Adviser to the Home Office

Dr Louise Slater
Associate Professor in Physical Geography, University of Oxford
Group Lead, Hydroclimate Extremes

2 Project Description

Tobler's first law of geography holds that "everything is related to everything else, but near things are more related than distant things." [Tob70]. This quirk of spatial data naturally poses challenges for traditional predictive modelling assumptions that observations are independent and identically distributed. Spatial forecasting is further complicated when data also has a temporal dimension and thus exhibits dependencies across both space and time.

Despite these challenges, spatio-temporal forecasting – making predictions about spatial data that changes over time – is a rapidly evolving field with increasing potential for applications in fields such as hydrology, public safety, and transportation, among many others [Xu+21; Yao+21; BVS22]. Common approaches employ methods from time series forecasting and spatial interpolation. However, the intersection between space and time present in these problems pose unique challenges that necessitate special consideration.

The spatio-temporal forecasting problem is posed as follows: take a set of locations $L = \{l_1, ..., l_n\}$, a set of time-stamps $T = \{t_1, ..., t_m\}$, and a set of observations

$$\mathcal{D} = \{\{y_{1,1}, (x^1_{1,1}, ..., x^k_{1,1})\}, ..., \{y_{i,j}, (x^1_{i,j}, ..., x^k_{i,j})\}\}^{j \in \{1,2,...,m\}}_{i \in \{1,2,...,n\}},$$

where $y_{i,j}$ and $x^k_{i,j}$ correspond, respectively, to the values of the output vector Y and the predictor matrix X_k at time t_i in geographical location l_j. [Oli21]. The objective is to predict the value of Y at a location of interest, $l_s, s \in \{1, 2, .., n\}$, at a time in the future, t_f, given the observed values $y_{i,j}$ and predictor vector $\mathbf{x_{i,j}}$ such that $t_m < t_f$.

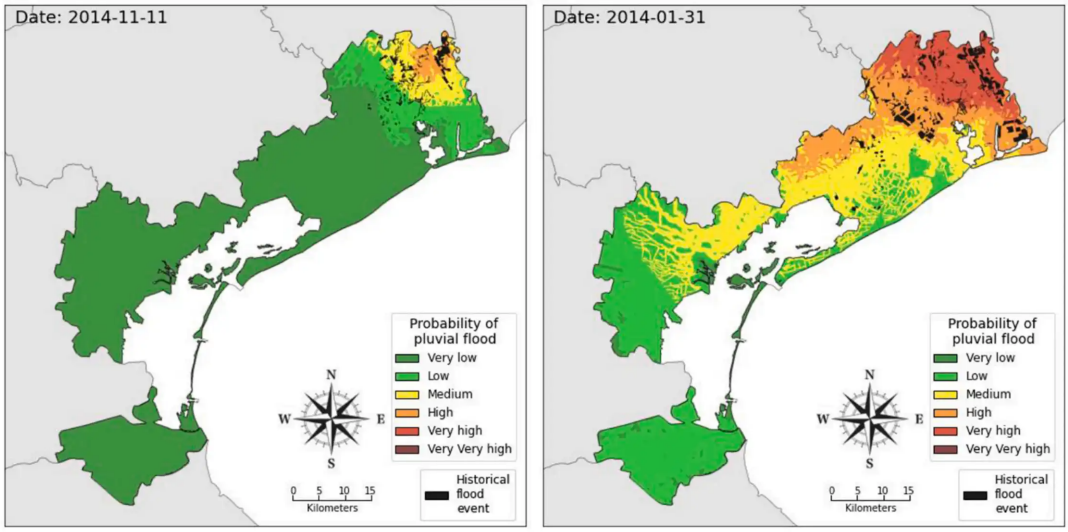

Figure 1: Example of a spatio-temporal forecast for pluvial flood areas in Venice. [Mar+22]

The most frequent predictive approach employed by spatio-temporal machine learning models consists of transforming the original problem into a multiple regression task, where the target variable is the future value of the series and the predictors are previous past values of the series up to a certain time window [OT12]. This transformation technique is known as time delay embedding. [Tak80] The goal is to provide the modelling algorithm with information on the recent dynamics of the time series by leveraging past predictors.

These approaches assume that the future conditions depend on recently observed conditions in the same location. However, certain spatio-temporal variables of interest (e.g. flooding, human trafficking, wind speeds) depend not only on the recent past conditions at the same location, but also on recent past conditions of nearby locations [OT12]. In other words, the "spatial radius of influence" may decrease for older observations such that observations that are more proximate in both time and location have a larger influence on the outcome variable than temporally and spatially distance observations. Thus, models that are feed only with values from the same location for which a future prediction is required may be limiting.

Ohashi and Torgo (2012) addressed this problem by proposing spatio-temporal indicators,

a method based on data pre-processing that uses statistics summarising past data within a spatio-temporal distance of the target observation as predictors [OT12]. These spatio-temporal indicators assume that data from distant neighbours becomes less relevant as the model looks further back in time. The main idea behind their proposal was to develop predictors capable of capturing the spatio-temporal dynamics of a given time series. These extra predictors then provide the model with important information on the recent spatio-temporal dynamics of the time series, which in turn improve the model prediction accuracy.

The model is specified as follows: the spatio-temporal indicators summarise values within spatio-temporal neighbourhoods of an observation, O, whose value is the target of the prediction [Oli21]. O is measured at time t_O and location l_O. Its spatio-temporal neighbourhood contains all past observations within specified boundary, β, of spatio-temporal distance

$$D_A = \alpha \cdot D_A^S + (1 - \alpha) \cdot D_A^T \leq \beta,$$

where D_A^S is the spatial distance between the locations where O and a reference point A occur, D_A^T is the temporal distance between observations O and A, α is a weighting factor that determines the relative importance of spatially vs. temporally proximate observations, and β is the chosen maximum spatio-temporal distance.

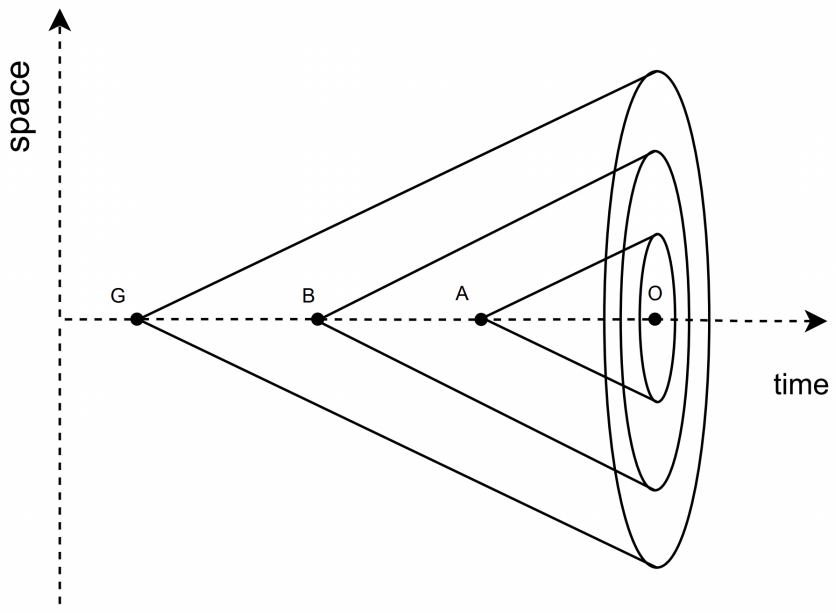

Figure 2: Spatio-temporal neighbourhoods of different sizes as defined by Ohashi and Torgo

The indicators are only based on past observations such that temporal distance is defined as

$$D_A^T = \begin{cases} \frac{t_O - t_A}{t_{max} - t_{min}}, & \text{if } t_A < t_O \\ \infty, & \text{otherwise.} \end{cases}$$

Note that the spatial and temporal distances D_A^S and D_A^T are normalized to be in the $[0, 1]$ interval such that the spatio-temporal boundaries are defined relatively, accounting for spatio-temporal distances differing within and between data sets.

Having defined the spatio-temporal distance between two observations, the spatio-temporal neighbourhood of an observation O can then be defined as the set of points within a certain spatio-temporal distance of that observation

$$N_O^\beta = \{A \in \mathcal{D} : D_A < \beta\},$$

where $\mathcal{D}$ is the available spatio-temporal dataset.

It is also true that spatio-temporal variables of interest may take some time to "travel" from one point to another [Oli21]. In this "reversed cone" paradigm, spatially distance observations have a larger effect on the predicted variable when they are further away temporally. For example, in order to predict flooding in a given region, it may be more useful to utilise observations from further upstream locations than immediately proximate ones. Similarly, human trafficking activity in a major city may be predictive of future human trafficking in smaller cities located further away after sufficient time has passed.

In this spatio-temporal relationship, the new temporal distance is defined as

$$D_A^{T\prime} = \begin{cases} \frac{\beta}{1-\alpha} - D_A^T, & \text{if } \frac{\beta}{1-\alpha} < D_A^T, t_A < t_O \\ \infty, & \text{otherwise.} \end{cases}$$

Thus the combined spatio-temporal distance is be recalculated as

$$D_A = \alpha \cdot D_A^S - (1-\alpha) \cdot D_A^T \leq 0.$$

These spatio-temporal indicators summarise the dynamics of the series within the neighbourhood. Given the above definitions, the spatio-temporal neighbourhood of a point can be thought of as a cone within space-time (Figure 2) [OT12]. Such cones represent which past values may influence the future value of the time series at that location and can be regarded as the spatio-temporal equivalents of time-delay embedding. Notably, different settings for α and β lead to cones of difference sizes and different definitions of temporal distance can change the orientation of the cone, altering the spatio-temporal embedded relationship.

Ohashi and Torgo (2012) as well as Oliveira (2021) combine spatio-temporal indicators with a wide range of modeling approaches (linear regression, multivariate adaptive regression splines, support vector machines, regression trees, and random forests) to greatly improve spatio-temporal predictions for specific data sets. However, the spatio-temporal relationships specified by the indicator equation are highly bespoke to the domain in which the forecast is produced. Thus changing parameters of the indicator greatly affect its utility as a predictor.

Thus far, there have been no efforts to determine the optimal spatio-temporal cone shape or orientation for a given data set. In order to generalize spatio-temporal indicators for broader applications, methods for selecting the optimal spatio-temporal indicator specification need to be developed. The proposed steps are the following:

1. Identify methods for optimizing parameters of spatio-temporal cone shape, α and β, as well as orientation, determined by D_A^T, for a given data set (e.g. producing a grid search algorithm or using numerical optimization, where possible)

2. Apply block cross-validation to evaluate parameter tuning methods on a variety of spatio-temporal data sets using different modeling approaches (linear regression, support vector machines, regression trees, random forests, convolutional neural networks)

3. If sufficient time remains, develop and publish a package in R to select optimal spatio-temporal indications for different spatial data types (raster, point, polygon, and line)

Methodology for selecting an optimal set of spatio-temporal indicators has the potential to substantially improve the quality of spatio-temporal forecasts. The applications of such indicators are uniquely relevant to dynamical models – where input values can be altered to study how outputs change virtually in real time – as spatio-temporal indicators can leverage real-time data collected from neighbouring locations to make predictions about the future [Sla+22; OT12]. The parameters that specify which spatio-temporal cone that is ultimately used in modelling also have an interpretable meaning for the subject area, enabling researchers to learn what spatio-temporal relations exist within their data (for example, which spatially distant predictors have a time delayed effect on the output).

3 Prerequisite Knowledge

3.1 Courses

- Applied Statistics (MT22)
- Statistical Inference (MT22)
- Statistical Programming (MT22)
- Algorithmic Foundations of Learning (MT22)
- Statistical Machine Learning (HT23)
- Advanced Topics in Statistical Machine Learning (HT23)

3.2 Tools

- R (`stars`, `gstat`, `STEvaluation`, `sp`, `raster`, `spacetime`, `rgdal`, `rgeos`, `earth`, `stats`, `forecast`, `e1071`, `randomForest`, `neuralnet`, `ggmap`, `ggplot2`)
- QGIS

4 Data Availability

Spatio-temporal data can be difficult to access given the density of information needed to produce such data. For the purposes of this dissertation, the publicly available atmospheric data sets used by Oliveira (2021) will be analysed.

Table 1: Spatio-temporal descriptions of real-world data sets to be used in analysis. [Oli21]

ID	Data source	ID	Variables	Time-stamps	Frequency	Nb. loc.	Network	Nb. obs.	Avail. (%)
1	MESA Air Pollution[1]	10	NO_X conc.	280	bi-weekly	20	irregular	5.6k	100
2	NCDC Air Climate[1]	20	precipitation	105	monthly	72	irregular	7.6k	100
		21	solar energy						
3	TCE Air Climate[1]	30	ozone conc.	330	hourly	26	irregular	8.6k	100
		31	air temperature	360				9.4k	
		32	wind speed	360				9.4k	
4	Cook Agronomy Farm[2]	40	water content	729	daily	40	irregular	22.3k	73
		41	temperature					22.5k	74
		42	conductivity					22.5k	74
5	SAC Air Climate[1]	50	air temperature	144	monthly	900	regular	130k	100
6	airBase[3]	60	PM10 conc.	4382	daily	70	irregular	149k	49
7	Beijing UrbanAir[4]	70	NO_X conc.	6.6k	hourly	36	irregular	152	64
		71	PM_{10} conc.					155k	66
		72	wind speed					161k	68
		73	PM_{25} conc.					162k	68
		74	humidity					162k	69
		75	air temperature					163k	69

[1] From Pravilovic et al. [2018]; Downloaded at: `http://www.di.uniba.it/appice/software/COSTK/data/dataset.zip`, accessed on 12 March 2018;
[2] Loaded from R package GSIF [Hengl, 2017, Gasch et al., 2015] version 0.5-5.1 (`https://cran.r-project.org/web/packages/GSIF/index.html`, accessed on 9 December 2020);
[3] Loaded from R package spacetime [Pebesma, 2012] version 1.2–3 (`https://cran.r-project.org/web/packages/spacetime/index.html`, accessed on 9 December 2020);
[4] From Zheng et al. [2013]; Downloaded at: `https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/Air20Quality20Data.zip`, accessed on 18 October 2017; Since there was more than one measurement for some hours we rounded the time-stamps to the closest hour, and calculated the median values per hour and location.

Additional data sets such as the UK's National River Flow Archive (accessible through the R package `rnrfa`) may also be examined [Sla+19]. It is also possible to analyse additional subject-matter relevant data sets under advisement from supervisors.

5 Computing Resources

All computing for this project will be completed on a personal laptop. Access to GPU(s) would be helpful, but not necessary.

References

[BVS22] Dhivya Bharathia, Lelitha Vanajakshia, and Shankar C. Subramanianb. "Spatio-temporal modelling and prediction of bus travel time using a higher-order traffic flow model." In: *Physica A: Statistical Mechanics and its Applications* 596 (2022). DOI: https://doi.org/10.1016/j.physa.2022.127086.

[Mar+22] Zanetti Marcoab et al. "Spatio-temporal cross-validation to predict pluvial flood events in the Metropolitan City of Venice." In: *Journal of Hydrology* 612 (2022). DOI: https://doi.org/10.1016/j.jhydrol.2022.128150.

[OT12] Orlando Ohashi and Luís Torgo. "Wind speed forecasting using spatio-temporal indicators." In: *Dynamical Systems and Turbulence, Warwick 1980* 898 (2012), pp. 366–381. DOI: doi:10.3233/978-1-61499-098-7-975.

[Oli21] Mariana Rafaela Figueiredo Ferreira de Oliveira. "Predictive Analytics for Spatio-Temporal Data." PhD thesis. Braga, Portugal: Universidades do Minho, Aveiro, e Porto, 2021.

[Sla+22] Louise Slater et al. "Hybrid forecasting: using statistics and machine learning to integrate predictions from dynamical models." 2022. DOI: https://doi.org/10.5194/hess-2022-334.

[Sla+19] Louise J. Slater et al. "Using R in hydrology: a review of recent developments and future directions." In: *Hydrology and Earth System Sciences* 23 (2019), pp. 2939–2963. DOI: doi:10.3233/978-1-61499-098-7-975.

[Tak80] Floris Takens. "Detecting strange attractors in turbulence." In: *Dynamical Systems and Turbulence, Warwick 1980* 898 (1980), pp. 366–381. DOI: https://doi.org/10.1007/BFb0091924.

[Tob70] W. R. Tobler. "A computer movie simulating urban growth in the Detroit region." In: *Economic Geography* 46 (1970), pp. 234–240. DOI: https://doi.org/10.2307/143141.

[Xu+21] Lei Xu et al. "Spatiotemporal forecasting in earth system science: Methods, uncertainties, predictability and future directions." In: *Earth-Science Reviews* 222 (2021). DOI: https://doi.org/10.1016/j.earscirev.2021.103828.

[Yao+21] Yao Yao et al. "Spatiotemporal distribution of human trafficking in China and predicting the locations of missing persons." In: *Computers, Environment and Urban Systems* 85 (2021). DOI: https://doi.org/10.1016/j.compenvurbsys.2020.101567.