Improved Spatial-Temporal Forecasting through Modelling of Spatial Residuals in Recent History

D. Pokrajac, and Z. Obradovic[*]

1 Introduction

Prediction of a continuous response variable in spatial-temporal domains has recently drawn attention in a data analysis community [4],[11]. Spatial-temporal regression models, learned on systematically collected values of driving attributes, can contribute to better understanding of complex phenomena studied in meteorology, oceanography, environmental science, precision agriculture and other domains. However, spatial-temporal modelling is often difficult due to various factors. For example, due to a prohibitive cost of data collection or other constraints, it is often not possible to systematically measure values of all attributes that have an influence to the observed response [15]. In such applications, models estimated on available attributes often have unsatisfactory explanatory power.

Possible solutions are auto-regressive models that use information from a spatial neighborhood to perform a prediction at a specified location. Performance improvements compared to ordinary regression models are often possible due to a postulated spatial correlation of data. Typical spatial prediction methods have been developed assuming non-uniform event-driven sampling [14], where the objective is

[*]Center for Information Science and Technology, Temple University, 303 Wachman Hall (038-24), 1805 N. Broad St. ,Philadelphia, PA 19122, USA,
e-mails: d_pokrajac@yahoo. com,zoran@joda. cis. temple. edu

interpolation at different spatial positions. Without modifications, these methods are not applicable to prediction of unknown future response values using uniform grid sampling.

Existing spatial-temporal prediction methods have difficulties to properly estimate a temporal part of the model when the number of available time layers is rather small. Also, a majority of spatial-temporal learning algorithms were developed for stationary or time-constant processes. Data non-stationarity can significantly decrease the prediction quality and the applicability of such prediction models.

Finally, for non-linear phenomena, learning algorithms with a response variable modelled as a non-linear function of driving attributes may be superior to linear predictors. However, due to the presence of noise in the data, insufficient size of a training set and interpolation error, linear models can in practice outperform non-linear ones [13].

The purpose of this study is to examine the effect of including auto-regressive modelling of ordinary regression error residuals for learning on spatial-temporal data sampled on a uniform grid. The proposed method combines linear or non-linear non-spatial and non-temporal regression models learned on data collected over time with spatial-temporal auto-regression of residuals.

After a survey of related work presented in Section 2, the proposed method for spatial-temporal prediction is described in Section 3. Experimental data properties, the accuracy measure for model evaluation and the obtained experimental results are reported in Section 4, followed by conclusions and directions for future work discussed in Section 5.

2 Related Work

2.1 Spatial Auto-regression

In analysis of spatial data, numerous attempts were made to explicitly include a spatial component into prediction models. In models with spatially correlated residuals and with auto-regressive disturbance ([10],[7]) modelling consists of two steps. First, the response variable is treated as non-spatial and a linear model is applied. Then, the residuals of a linear model on training data are assumed to be spatially correlated and their dependence is modelled through a "neighborhood" matrix using an auto-regressive approach.

In these models, the objective is to estimate the response y at a desired location as a function of k attributes observed at that location and of the prediction errors $\mathbf{u} = [u_1, \ldots, u_n]^T$ at given n training examples. So, a training set of consists of n patterns $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ with $\mathbf{x}_i = [x_{i,1} \ldots x_{i,k}]^T$ being k observed attributes at a specific location in space and y_i the corresponding response. In a matrix representation, these models can be described as:

$$\begin{aligned} \mathbf{y} &= X\boldsymbol{\beta} + \mathbf{u} \\ \mathbf{u} &= W\mathbf{u} + \boldsymbol{\epsilon} \end{aligned} \tag{1}$$

Here, $\boldsymbol{\beta} = [\beta_0 \beta_1 \ldots \beta_k]^T$ is vector of parameters for an ordinary linear regression model, $X = \begin{bmatrix} 1 & x_{1,1} & \dots & x_{1,k} \\ 1 & \dots & \dots & \dots \\ 1 & x_{n,1} & \dots & x_{n,k} \end{bmatrix}$ is $n \cdot (k+1)$ matrix containing a unit $n \cdot 1$ vector and n vectors of k observed attributes, $\boldsymbol{\epsilon} = [\epsilon_1 \dots \epsilon_n]^T$ is a vector of independent identically distributed Gaussian disturbances and $W = [w_{i,j}]_{n,n}$ is an $n \cdot n$ neighborhood matrix having zeros on the main diagonal and other, pre-specified, positions.

An alternative representation of (1) is:

$$\mathbf{y} = X\boldsymbol{\beta} + W(\mathbf{y} - X\boldsymbol{\beta}) + \boldsymbol{\epsilon} \tag{2}$$

Models (1)-(2) can be estimated using a generalized least squares method or maximum likelihood techniques [14] depending whether values of non-zero elements of W are known in advance.

Another promising approach is the mixed regressive-spatial auto-regressive model [7], which is a generalization of models proposed in [2],[7]. This model assumes a spatially correlated response variable also dependent on attribute values at the neighboring points. Spatial dependence is specified using a column vector $\boldsymbol{\gamma} = [\gamma_1 \dots \gamma_k]^T$ of cross-correlation coefficients between attributes, an $n \cdot n$ neighborhood matrix W and a proportionality coefficient ρ:

$$\mathbf{y} = X\boldsymbol{\beta} + WX\boldsymbol{\gamma} + \rho W\mathbf{y} + \boldsymbol{\epsilon} \tag{3}$$

Models (2) and (3) as well as other spatial regression models introduced in the literature [2],[7] have the following common properties:

- Models are applicable for interpolation of non-uniform event-driven samples and not for prediction on an unseen data layer;
- For a response variable, data-generation process (DGP) constant in time is assumed;
- The response variable is assumed to be linearly dependent on driving attributes.

2.2 Temporal Auto-Regression

Temporal data can be modelled using a serial-correlation model [5] where the response variable is assumed to be a function of driving attributes, while residuals are assumed to be serially correlated, satisfying AR(1) model [1]. Let $\mathbf{x}_t = [1\; x_{1,t} \dots x_{k,t}]^T$ contain k attribute values at time instant t in addition to a constant 1 due to an additive constant factor β_0 in a linear model. Let y_t, u_t and ϵ_t indicate the response value, a correlated residual and a value of a white Gaussian noise at time instant t, respectively. Then, serial-correlation of residuals is modeled using the autocorrelation coefficient ρ:

$$\begin{aligned} y_t &= \beta_0 + \sum_i x_{i,t}\beta_i + u_t \\ u_t &= \rho \cdot u_{t-1} + \epsilon_t \end{aligned} \tag{4}$$

To estimate a serial-correlation model, one can perform the following iterative procedure [5]:

1. Set autocorrelation coefficient $\rho = 1$;
2. Assuming ρ is a constant, estimate ordinary regression coefficients $\boldsymbol{\beta}$ in the linear model:

$$y_t - \rho y_{t-1} = \beta_0(1-\rho) + \sum_i (x_{i,t} - \rho x_{i,t-1})\beta_i + u_t \qquad (5)$$

3. Estimate ordinary regression residuals $\hat{u}_t$ and $\hat{u}_{t-1}$;
4. Compute a value of ρ for the next iteration by estimating the regression model:

$$\hat{u}_t = \rho \hat{u}_{t-1} + \epsilon_t$$

5. Repeat steps 2-4 until a pre-specified convergence criterion is satisfied.

In this model, the stationarity of regression coefficients $\boldsymbol{\beta}$ is implicitly assumed, which is equivalent to imposing a modelling restriction $\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1}$. To properly estimate a serial-correlation model, the existence of a relatively high number of data time layers is necessary. Recall that these are strong and often unattainable requirements for a number of spatial-temporal domains (e.g. in precision agriculture, due to a recent adoption of a global positioning system-based measurement technology, spatial data currently exist for about 5 years). Also, observe that the serial correlation model does not consider a spatial correlation of the data. Therefore, a new class of spatial-temporal models has recently been developed.

2.3 Spatial-temporal Modelling

Spatial-temporal prediction can be performed using a generalization of the model with auto-regressive disturbance, defined in equation (2). Here, a residual neighborhood matrix W represents spatial-temporal correlation of residuals. This matrix is estimated assuming that second-order statistics of residuals satisfy theoretical spatial-temporal variograms [4]. Modelling includes estimation of linear regression coefficients $\boldsymbol{\beta}$ and computation of residuals, as well as the estimation of spatial-temporal variograms. When the model is estimated, prediction is performed using a weighted sum of driving attributes and residual estimation obtained by spatial-temporal kriging. Similar to serial correlation models, here regression coefficients are constant in time and a relatively large number of temporal observations is necessary for proper parameter estimation.

Another approach for regression of spatial-temporal data, proposed at [11], is a generalization of the mixed regressive-spatial auto-regressive model, defined at equation (3). Here, the neighborhood matrix is assumed to be a product of matrices T and S related to time and space dependence, respectively. Each sample from the training data is assumed to be dependent on a fixed number of spatial neighbors

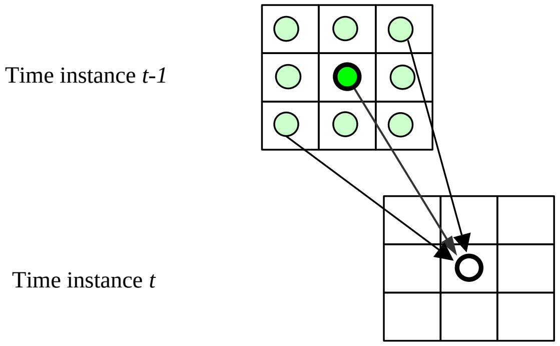


Figure 1. *An illustration of the proposed method: Information from a previous time instance at a desired location sample and its neighborhood is used for predicting current time response at this location.*

(regardless the time) and a fixed number of time neighbors (that immediately precede the observed sample). The main problem when applying this model is the correct estimation of T and S. The model is developed for real-estate data, where each sample occurs in a distinct time instant and the number of time neighbors considered is small. In contrast, in the case of a uniform grid considered in our study there are a larger number of samples collected at each time moment. Hence, the resulting matrix T is huge and the application of this method can be prohibitively laborious.

3 Methodology

In this paper, we propose a model with spatially correlated lagged residuals. The proposed model is a generalization of the serial correlation model [5] that includes residuals from the spatial neighborhood of an observed data sample. By this means, spatial-temporal correlation of ordinary regression residuals can be exploited.

At time layer t, in the i-th sampling point, residual $u_{t,i}$ of a (linear or nonlinear) regression model $y_\alpha(\mathbf{x}_{t,i})$ with parameters $\boldsymbol{\alpha} = [\alpha_1 \ldots \alpha_m]^T$ and the corresponding attribute vector $\mathbf{x}_{t,i}$, is assumed dependent on white Gaussian noise $\epsilon_{t,i}$ and residuals from the same sampling point and its neighbors at the previous time instant, as illustrated at Fig.1. The dependence between residuals at the i-th example at time instant t and the j-th example at time instant $t-1$ is described by the coefficient $w_{i,j}$ of the neighborhood matrix W.

The proposed model can be described as:

$$\begin{aligned} y_{t,i} &= y_\alpha(\mathbf{x}_{t,i}) + u_{t,i},\ i = 1, \ldots, n \\ \mathbf{u}_t &= W\mathbf{u}_{t-1} + \boldsymbol{\epsilon}_t \end{aligned} \tag{6}$$

where

$$\mathbf{u}_t \equiv [u_{t,1} u_{t,2} \ldots u_{t,n}]^T, \boldsymbol{\epsilon}_t \equiv [\epsilon_{t,1} \epsilon_{t,2} \ldots \epsilon_{t,n}]^T$$

Parameters of the proposed model can be estimated using minimization of the sum of squared errors in time instant t. After parameter estimation, model is tested on data from time instant $t+1$ using driving attributes collected in time instant $t+1$ and residuals computed using response values and their predictions at time t. Model coefficients are then re-estimated for the next time layer.

For pre-specified parameter values W and $\boldsymbol{\alpha}$, the response prediction at time t is computed as:

$$\hat{y}_{t,i} = y_\alpha(\mathbf{x}_{t,i}) + W[y_{t-1,1} - y_\alpha(\mathbf{x}_{t-1,1}) \ldots y_{t-1,n} - y_\alpha(\mathbf{x}_{t-1,n})]^T,\ i = 1, \ldots, n \quad (7)$$

Using an asymptotic analysis [5], it can be shown that parameter estimation by minimization of the sum of squared errors

$$SSR_t = \sum_{i=1,n} (\hat{y}_{t,i} - y_{t,i})^2 \quad (8)$$

is consistent (true values and expectations of their estimates are equal) if the stochastic process $\mathbf{u}_t = W\mathbf{u}_{t-1} + \boldsymbol{\epsilon}_t$, is stationary. The sufficient condition for this stationarity is that the matrix power series $\sum_{i=1,\infty} W^i$ converges, which is satisfied when [9]:

$$\max_i \sum_{j=1,n} |w_{i,j}| < 1 \quad (9)$$

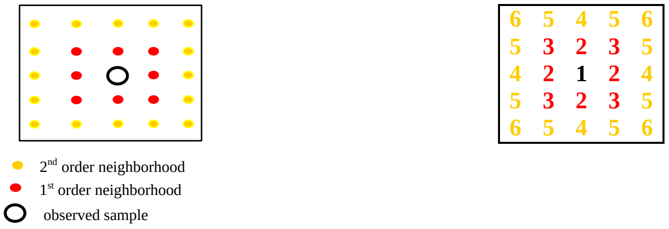
We impose the following restrictions to the model in order to limit the influence of residuals to spatial neighborhoods of the example and to impose spatial stationarity and isotropy:

- $w_{i,j} = 0$ unless two corresponding spatial samples are at most L-th order neighbors.

  Here, two examples are called the l-th order neighbors if maximal absolute difference of their spatial coordinates is $l\Delta$, where Δ is a sampling distance (see an example at Fig.2). The maximal order L of neighbors is an input parameter of the algorithm. For $L = 0$, prediction is performed using non-spatial residuals (only a past residual at the same position). In that case, the proposed model is equivalent to the serial-correlation model (4), but here estimated on multiple realizations of time-series, each corresponding to an example.

- $w_{i,j}$ depends only on the distance between corresponding examples regardless of their position at the spatial layer.

  Due to spatial stationarity, each row of matrix W will consist of the same, but permuted, values. Also, due to the imposed isotropy, radial symmetric neighbors have the same coefficient values in W, which reduces the number of relevant coefficient values in each row of W to $\nu \equiv (L+1) \cdot (L+2)/2$ (For $L = 1,2$ this can be easily confirmed on Fig.2b).



6	5	4	5	6
5	3	2	3	5
4	2	1	2	4
5	3	2	3	5
6	5	4	5	6

a) Definition of a neighbor order for spatial data with a uniform grid

b) The examples radial symmetric to the observed sample

Figure 2. *Spatial neighborhoods of an example.*

Due to imposed restrictions, model (6) reduces to:

$$y_{t,i} = y_\alpha(\mathbf{x}_{t,i}) + \tilde{w}_t^T(\tilde{\mathbf{y}}_{t-1,i} - \tilde{\mathbf{y}}_\alpha(\mathbf{x}_{t-1,i})) + \epsilon_{t,i}, i = 1, \ldots, n \qquad (10)$$

where ν-dimensional vectors $\tilde{\mathbf{y}}_{t-1,i}$ and $\tilde{\mathbf{y}}_\alpha(\mathbf{x}_{t-1,i})$ contain response values and ordinary regression prediction values corresponding to neighbors up to the order L at the previous time layer. Neighborhood dependence in (10) is modelled with a ν-dimensional vector $\tilde{\mathbf{w}}_t$. The first element of each vector corresponds to the i-th sample at time layer $t-1$. Due to assumed isotropy, samples on positions radial symmetric to the current sample should have equal weights. After an ordinary regression model is evaluated, vectors $\tilde{\mathbf{y}}_{t-1,i}$ and $\tilde{\mathbf{y}}_\alpha(\mathbf{x}_{t-1,i})$ contain sums of responses/ordinary regression predictions on radial symmetric positions. For example, when $L = 2$ vector $\tilde{\mathbf{y}}_{t-1,i}$ contains $\nu = 6$ values: the response at the observed position and sums of responses at positions bearing equal numbers on Fig.2b (each number corresponds to a distinct element of $\tilde{\mathbf{y}}_{t-1,i}$, $\tilde{\mathbf{y}}_\alpha(\mathbf{x}_{t-1,i})$and $\tilde{\mathbf{w}}_t$) .

The stationarity condition (9) implies that least-squares estimation of (10) is consistent when

$$\sum_i |\tilde{w}_{t,i}| < 1 \qquad (11)$$

In the rest of the paper, this consistency will be assumed.

Observe that model (10) is non-linear, since $\tilde{\mathbf{w}}_t$ multiplies $\tilde{\mathbf{y}}_\alpha(\cdot)$. Hence, parameters $\tilde{\mathbf{w}}_t$ and $\boldsymbol{\alpha}$ can be estimated using standard methods for non-linear optimization [5]. However, when an ordinary regression component $y_\alpha(\cdot)$ is non-linear, such methods become computationally expensive, so the following procedure can be applied:

- Estimate non-spatial ordinary regression models on data from time layers $t-1$ and t and compute the corresponding estimated residuals $\tilde{\mathbf{u}}_t$,$\tilde{\mathbf{u}}_{t-1}$;
- Perform spatial autoregression of $\tilde{\mathbf{u}}_t$ on $\tilde{\mathbf{u}}_{t-1}$ and estimate $\tilde{\mathbf{w}}_t$.

Observe that this technique is similar to the first iteration of Cochrane-Orcutt algorithm [5] that takes into consideration lagged neighboring residuals. However here, due to instability of the initial estimation $\tilde{\mathbf{w}}_t$ (because of errors in the computation of estimated residuals), further iterations in an algorithm analogous to that of Cochrane-Orcutt cannot result with useful estimation.

Unlike a serial correlation model which is time-series oriented, the proposed method relies on spatially organized data and can exploit the fact that the number of spatial sample points is large compared to the number of observed time instants.

Similar to spatial-temporal auto-regression [11], the proposed model considers the influence of time and space neighbors separately. Further, in both models the maximum size of a spatial neighborhood influence must be pre-specified. However, in contrast to the spatial-temporal auto-regressive model, the proposed model does not involve a spatial regression on attributes. Also, in the proposed model, the maximal time lag of considered residuals is one, which makes prediction less time-complex and potentially more resistant to data non-stationarity.

Residuals in the proposed model are correlated with residuals in neighboring points, similar to disturbance in spatial auto-regression models [2],[7],[10]. However, in the proposed model this correlation is established with residuals in the time layer prior to the predicted value, making the prediction of response values in the future feasible.

The proposed method is similar to a spatial-temporal auto-regressive model proposed in [16], in sense that both methods employ spatial-temporal correlation of data to improve prediction accuracy. However, in contrast to the former approach, the method we propose performs auto-regression on residuals of the attribute modeling rather than directly on the response variable.

Compared to generalization of the model with auto-regressive disturbance [4], the proposed model has more degrees of freedom and therefore potentially higher explanatory power.

4 Experiments

4.1 Properties of Experimental Data

Experiments were performed on data generated using our spatial-temporal data simulator [12]. Data with controllable complexity were generated to satisfy pre-specified spatial and temporal statistical properties. Simulated agricultural data consisted of five time layers. Data contained samples of 5 simulated driving attributes and the response variable. Each time layer consisted of $n = 6561$ samples from a $80 * 80m^2$ rectangular field, with $10m$ sampling distance. The mean and standard deviation of the simulated crop yield were similar to that of real-life data.

Driving attributes were generated through a multistep process of grid determination, generation of spatially correlated attributes and cluster generation [12]. For each attribute, time layers were generated using kriging of a random seed vector [3]. Using Cholesky decomposition, a seed vector was generated to satisfy specified spatial and temporal correlation ([12]; Pokrajac, Obradovic, unpublished results).

Five simulated attributes had a spatial correlation similar to the following real-life agricultural variables: nitrogen (N), phosphorus (P), potassium (K), profile curvature (C) and slope (S). Attributes C and S were assumed constant in time,

Table 1. *Spatial and temporal statistic parameters of simulated driving attributes.*

Attribute name		N	P	K	C	S
Spatial parameters	Range(m)	200	300	400	100	200
	Nugget(%)	0	0	0	0	0
Temporal parameters	% temporal variability	80	20	10	attributes do not change over time	
	Correlation	0.9	0.9	0.7		

while the remaining attributes were modelled as time-dependent. In the absence of real-life data temporal statistics, percentage of total variability due to temporal variance was varied in the range $10 - 80\%$, while the auto-correlation of successive time layers was chosen according to expert estimation (see Table 1).

After the generation of correlated attributes, four clusters in the space of topographic attributes C and S were formed. To generate attribute clusters, corresponding cluster "seeds" were chosen and each data point was "moved" towards the nearest cluster seed. The intensity of the shift proportional to the distance to the seed point was adjusted to control cluster aggregation. "Perturbation" noise with variances proportional to that of the attributes was introduced to avoid unnaturally clear separation between clusters. Crop yield, the response variable, was generated using two different models to simulate linear and non-linear data generation process (DGP).

A linear DGP was simulated using linear model to generate crop yield as a linear combination of attribute values. Non-linear DGP was simulated using linear plateau models (particularly suitable for agriculture applications) where the response variable was the product of linear plateau functions corresponding to each driving attribute. Here, linear plateau functions were assumed constant in time. The relative influence of particular attributes on the simulated response and the shape (slope and thresholds) of linear plateau functions were varied according to expert knowledge. Both homogeneous and heterogeneous DGP were simulated. In the homogeneous case, a single model parameterization was used to generate the response in all simulated data points while the response for a heterogeneity scenario was simulated by applying a separate model to each attribute cluster.

To investigate temporal variability of the response variable influenced by an external factor, an additional temporal component of a response, unexplainable by driving attributes, was simulated with AR(1) models [1]. An AR(1) process with autocorrelation coefficient 0.5, that comprised 10% of the total variability of simulated response, was applied using both "cluster-wise" and "point-wise" assignment methods. In a "cluster-wise" assignment, one realization of an AR(1) process was assigned to each attribute cluster, consequently modelling a real-life situation that particular zones of a spatial area behave differently in time. In contrast, in a "point-wise" assignment, one independent realization of a specified AR(1) process was generated per each simulated spatial point.

4.2 Accuracy Evaluation

Before regression was performed, data were normalized such that driving attributes and the response variable have zero mean and unit standard deviation. Linear regression was performed using the OLS method [6]. Non-linear modelling was performed by feedforward neural networks with 1 hidden layer, containing 4 neurons, and sigmoidal activation. Networks were trained using the Levenberg-Marquardt algorithm [8]. Since our primary intention was not to achieve an "optimal" non-linear model, further optimization of neural network topology and training algorithm was not performed.

The prediction accuracy at a time instance t was estimated by measuring explained response variability using the coefficient of determination R_t^2, defined as:

$$\begin{aligned} R_t^2 &= \frac{SSR_t}{SST_t} \\ &\text{where} \\ SST_t &= \sum_{i=1,n} (y_i - \bar{y})^2, \bar{y} = \frac{1}{n} \sum_{i=1,n} y_i \end{aligned} \tag{12}$$

For useful prediction models R^2 ranges from 0 to 1, with larger scores obtained by more accurate predictors, where 0 score corresponds to using a trivial mean predictor and 1 represents the ideal case of no prediction error. To obtain a correct estimate of non-linear model accuracy, experiments with non-linear models were repeated 10 times each and average R^2 values were reported. The training of the proposed method required data from two successive layers, and therefore trained models were tested on data from time layers 3,4 and 5 of five generated layers.

The proposed method that takes into consideration spatial residuals ($L > 0$) was compared to ordinary regression models and a serial-correlation model (where $L = 0$). Since these models are special cases of the proposed model with constraints, a likelihood ratio-type test [5] was applied. To perform the test, SSR values for the proposed (unconstrained) and an alternative model (constrained) were evaluated and for each time layer the ratio:

$$lr = \frac{SSR_{unconstrained} - SSR_{constrained}}{SSR_{unconstrained}/(n-p)} \tag{13}$$

was calculated, where p denotes the number of parameters in an unconstrained model. According to asymptotic theory [5], the lr ratio has an asymptotically χ^2 distribution, under a null hypothesis that there is no significant difference in performance due to releasing of constraints (an introduction of the proposed instead of an alternative model). The degrees of freedom for χ^2 distribution is equal to $\nu - 1$, when proposed model was compared to a serial-correlation model or ν, when compared to the ordinary regression models. Here, ν is the dimension of vectors $\tilde{\mathbf{y}}_{t-1,i}$ and $\tilde{\mathbf{y}}_\alpha(\mathbf{x}_{t-1,i})$ from equation (10). The null hypothesis was rejected whenever lr ratio was larger than the 99.9% quantile of a χ^2 distribution, ensuring 99.9% confidence in the rejection.

Spatial correlation of ordinary regression residuals was estimated using a robust estimation of spatial semi-variograms [3].

4.3 Results

Prediction results on simulated data are shown in Tables 2-5. In each table, the prediction accuracy is shown for three models of temporal dependence of the response and for two sets of attributes used for prediction. For each temporal dependence/attribute set combination, an ordinary regression model, along with the proposed method for no-spatial ($L = 0$) and spatial ($L > 0$) temporal residual regression, is evaluated for three time layers. For each time layer, a statistically significant improvement of the non-spatial model vs. an ordinary regression model was denoted with †. Asterisk (*) was used to denote a significant improvement due to the application of the proposed method for $L > 0$.

Prediction results on data generated using linear data generation process (DGP) are shown in Table 2. Assuming access to all driving attributes and no temporal variability of the response, a linear model was capable of explaining the complete variance of response ($R^2 = 1$). The presence of temporal variability decreased the prediction accuracy. When the proposed method with $L = 0$ was applied, significant performance improvement compared to an ordinary regression model was achieved with AR(1) disturbance added point-wise to the response. However, with a "cluster-wise" temporal disturbance (where a realization of a AR(1) process was assigned to each cluster), the improvement was achieved in only two of three time layers (time instants 3 and 5). Since in the "cluster-wise" case the number of actual realization of AR(1) process was equal to the number of clusters, the proposed model could not predict correctly parameters of temporal dependence. However, with a large number of AR(1) realizations (as in the point-wise case where the number of realization was equal to the number of points), the proposed method correctly estimated the correlation of AR(1) process, which resulted in improved accuracy. Since temporal dependency was not spatially correlated, the inclusion of neighbors could not improve the prediction: hence, there is no further increase of R^2 when the proposed method was applied with $L = 1$. With temporal attributes missing (not used for training and prediction), residuals of ordinary regression were spatially and temporally correlated (due to DGP linearity in missing temporal attributes). The proposed method with $L = 0$ improved accuracy when there was no temporal disturbance, or when such disturbance could be correctly estimated based on available samples (i.e. when AR(1) disturbance was applied point-wise). In these cases, due to spatial correlation of ordinary regression residuals, significant improvements were achieved when spatial residuals were introduced in prediction model ($L = 1$). No further improvements were achieved with $L = 2$ (these results are omitted from Table 2).

When data were simulated using a homogeneous non-linear DGP, we wanted to confirm a spatial-temporal structure of residuals prior to testing the proposed method that would exploit such a structure. Indeed, ordinary linear regression residuals were spatially correlated, as illustrated on Fig.3 for time layer 4. Compared with an approximately flat normalized semivariogram of a spatially uncorrelated noise, the semivariograms of residuals for all time layers significantly increase with distance, Fig.4, thus confirming spatial correlation of residuals. Spatial-temporal correlation was verified modeling residuals at $t + 1$ by linear regression on residuals

Table 2. *Comparison of the proposed method and ordinary linear regression. Experiments were performed on simulated data generated using a linear data generation process (DGP).*

TEMPORAL DISTURBANCE OF THE RESPONSE	TIME LAYER	PREDICTION ACCURACY (R^2)					
		No missing attributes			Temporal attributes (N,P,K) missing		
		Ordinary regression by OLS	Proposed method		Ordinary regression by OLS	Proposed method	
			L=0	L=1		L=0	L=1
No temporal variability	3	1	1	1	0.31	$0.60^\dagger$	0.64*
	4	1	1	1	0.66	$0.89^\dagger$	0.90*
	5	1	1	1	0.66	$0.92^\dagger$	0.95*
Clusterwise AR(1)	3	0.70	$0.74^\dagger$	0.74	< 0	< 0	< 0
	4	0.88	0.60	0.60	0.60	0.37	0.34
	5	0.44	$0.71^\dagger$	0.71	0.21	$0.62^\dagger$	0.66*
Pointwise AR(1)	3	0.89	$0.92^\dagger$	0.92	0.28	$0.56^\dagger$	0.59*
	4	0.89	$0.92^\dagger$	0.92	0.31	$0.60^\dagger$	0.64*
	5	0.89	$0.92^\dagger$	0.92	0.66	$0.89^\dagger$	0.90^*

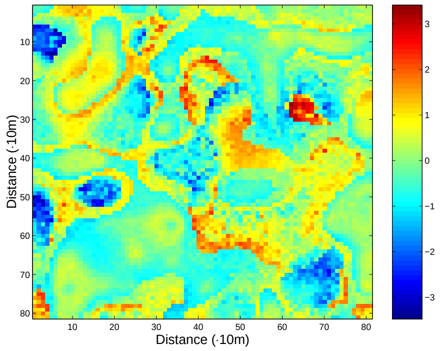


Figure 3. *Spatial placement of ordinary linear regression residuals for temporal layer $t = 4$ on data generated with non-linear homogeneous DGP.*

at $t = 1, 2, 3, 4$. Using a LR-type test, a significant (confidence 99.9%) improvement in prediction accuracy was shown when prediction was performed using a spatial neighborhood of each residual.

The prediction results of a linear model on data simulated by a homogeneous non-linear DGP are shown in Table 3. Using a non-spatial variant of the proposed method ($L = 0$), significant improvements were achieved regardless of the presence and type of temporal dependence at the response. As earlier, improvements were larger when temporal driving attributes were unobserved. When all driving attributes were available for training a prediction model (no missing attributes), the application of the proposed method using neighboring residuals (with $L = 1$) led

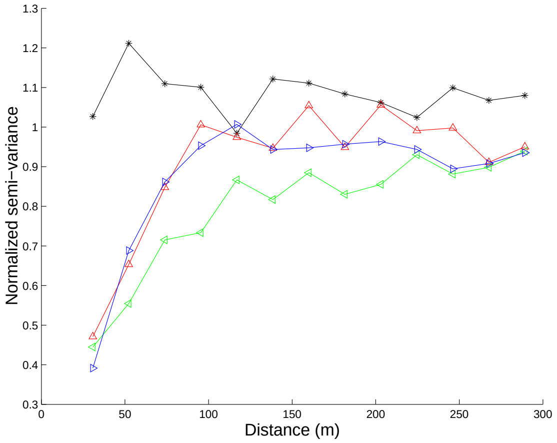


Figure 4. *Semi-variogram comparison of simulated uncorrelated noise (-∗-) and ordinary linear regression residuals on data generated with non-linear homogeneous DGP for three time instants: t=3 (-▷-),t=4 (-Δ-) and t=5 (-◁-).*

Table 3. *Comparison of the proposed method and ordinary linear regression. Experiments were performed on simulated data generated using a homogeneous linear data generation process (DGP).*

TEMPORAL DISTURBANCE OF THE RESPONSE	TIME LAYER	PREDICTION ACCURACY (R^2)					
		No missing attributes			Temporal attributes missing		
		Ordinary regression by OLS	Proposed method		Ordinary regression by OLS	Proposed method	
			L=0	L=1		L=0	L=1
No temporal variability	3	0.79	$0.90^{\dagger}$	0.91^{*}	0.20	$0.53^{\dagger}$	0.59^{*}
	4	0.83	$0.91^{\dagger}$	0.86	0.52	$0.54^{\dagger}$	0.55^{*}
	5	0.77	$0.88^{\dagger}$	0.89^{*}	0.30	$0.70^{\dagger}$	0.74^{*}
Clusterwise AR(1)	3	0.68	$0.84^{\dagger}$	0.87^{*}	< 0	$0.21^{\dagger}$	0.35
	4	< 0	< 0	< 0	< 0	< 0	< 0
	5	0.77	$0.77^{\dagger}$	0.71	0.33	$0.73^{\dagger}$	0.58
Pointwise AR(1)	3	0.64	$0.77^{\dagger}$	0.78^{*}	0.16	$0.45^{\dagger}$	0.50^{*}
	4	0.69	$0.81^{\dagger}$	0.83^{*}	0.44	$0.72^{\dagger}$	< 0
	5	0.62	$0.76^{\dagger}$	0.77^{*}	0.24	$0.59^{\dagger}$	0.64^{*}

to the significant improvement of R^2 when the response variable had a point-wise temporal component. Further increment of the neighborhood size ($L = 2, 3 \ldots$) did not result in a significant accuracy improvement (hence in Tables 3-5 only results for $L = 0, 1$ were reported).

When temporal attributes were missing, the prediction accuracy was improved using the proposed method with $L = 1$ if no additional temporal dependence of response was present. The lack of improvement of $L = 1$ vs. $L = 0$ in other

Table 4. *Comparison of the proposed method and ordinary linear regression. Experiments were performed on simulated data generated using a heterogeneous linear data generation process (DGP).*

TEMPORAL DISTURBANCE OF THE RESPONSE	TIME LAYER	PREDICTION ACCURACY (R^2)					
		No missing attributes			Temporal attributes missing		
		Ordinary regression by OLS	Proposed method		Ordinary regression by OLS	Proposed method	
			L=0	L=1		L=0	L=1
No temporal variability	3	0.24	$0.80^{\dagger}$	0.82^{*}	< 0	$0.28^{\dagger}$	0.30^{*}
	4	0.24	$0.82^{\dagger}$	0.83^{*}	0.07	$0.71^{\dagger}$	0.57
	5	0.12	$0.70^{\dagger}$	0.72^{*}	< 0	$0.61^{\dagger}$	0.64^{*}
Clusterwise AR(1)	3	0.26	$0.35^{\dagger}$	0.35	0.30	< 0	< 0
	4	0.16	< 0	0.44^{*}	< 0	< 0	0.37^{*}
	5	0.05	$0.69^{\dagger}$	0.69	< 0	$0.63^{\dagger}$	0.64^{*}
Pointwise AR(1)	3	0.18	$0.60^{\dagger}$	0.65^{*}	< 0	$0.34^{\dagger}$	0.47^{*}
	4	0.20	$0.70^{\dagger}$	0.73^{*}	0.06	$0.47^{\dagger}$	0.58^{*}
	5	0.11	$0.54^{\dagger}$	0.56^{*}	< 0	$0.80^{\dagger}$	0.49^{*}

examined cases is owing to the fact that the spatial structure of residuals did not reflect the spatial correlation of missing attributes (as it would have if an additive plateau model had been used instead of a multiplicative plateau model). In addition to cases where improvements were achieved for $L = 1$, the proposed method with $L = 0$ improved the accuracy when used on data with the point-wise temporal dependence. On the other hand, in the case of a cluster-wise AR(1) model, temporal dependence was rather difficult to have been correctly estimated, so the proposed method could not provide a consistent performance improvement.

The results of experiments on data with heterogeneous DGP are shown in Table 4. Here, due to the existence of additional spatial structure in data (each simulation model was assigned to one data cluster), the proposed model with $L = 1$ led to significant performance improvements more consistently. With a point-wise temporal response dependence and missing temporal attributes, the R^2 improvement was huge. Also, the improvement was significant when all attributes were used for prediction, if the additional temporal dependence did not exist or could be correctly predicted (point-wise model). Due to data heterogeneity, the proposed method with $L = 1$ typically had worse performance at points on the cluster boundaries. As in previous experiments, with a cluster-wise temporal disturbance, results of the proposed method using $L = 1$ were not always satisfactory. On the other side, a significant accuracy improvement using the proposed method with $L = 0$ was achieved whenever temporal dependence of residuals had been properly discovered.

Results of experiments on heterogeneous data repeated with neural networks as non-linear ordinary regression models are shown on Table 5. Comparing Tables 4 and 5, one can conclude that replacement of linear models with neural networks resulted in significant accuracy improvement when an ordinary regression is performed. However, accuracy of linear regression models when combined with the proposed method was considerably higher than the accuracy of ordinary non-linear regression models.

Table 5. *Comparison of the proposed method and ordinary regression when neural networks were used as a non-spatial and non-temporal model. Experiments were performed on simulated data generated using heterogeneous (cluster-specific) plateau DGP.*

TEMPORAL DISTURBANCE RESPONSE	TIME LAYER	PREDICTION ACCURACY (R^2)					
		No missing attributes			Temporal attributes missing		
		Ordinary regression by neural networks	Proposed method		Ordinary regression by neural networks	Proposed method	
			L=0	L=1		L=0	L=1
No temporal variability	3	0.60	$0.77^\dagger$	0.80^*	< 0	$0.28^\dagger$	0.32^*
	4	0.62	$0.91^\dagger$	0.90	0.35	$0.83^\dagger$	0.86^*
	5	0.60	$0.78^\dagger$	0.80^*	0.17	$0.64^\dagger$	0.66^*
Clusterwise AR(1)	3	0.42	$0.51^\dagger$	0.48	< 0	$0.01^\dagger$	0.03^*
	4	0.52	$0.73^\dagger$	0.75^*	0.21	$0.68^\dagger$	0.71^*
	5	0.52	$0.76^\dagger$	0.75	0.07	$0.56^\dagger$	0.58^*
Pointwise AR(1)	3	0.42	$0.61^\dagger$	0.65^*	0.02	$0.34^\dagger$	0.39^*
	4	0.50	$0.74^\dagger$	0.77^*	0.31	$0.69^\dagger$	0.73^*
	5	0.51	$0.67^\dagger$	0.67	0.12	$0.48^\dagger$	0.53^*

When temporal attributes were missing, the proposed method with spatial residuals ($L = 1$) significantly outperformed an ordinary neural network regression model. With no missing attributes, the proposed method with $L = 0$ had a significant improvement vs. ordinary regression (left half of Table 5), and further improvements using $L = 1$ still existed, but were not consistent.

5 Conclusions and Further Research

In this study, a spatial-temporal data prediction technique based on a combination of non-spatial ordinary regression and spatial-temporal auto-regression of residuals is proposed. Using simulated spatial-temporal data of various complexity, the proposed method was compared to ordinary linear and non-linear models. Experimental results suggest that the proposed method can significantly improve the prediction accuracy of linear regression models and make their accuracy similar to or better than the accuracy of ordinary non-linear models. Further accuracy improvement is possible when the method is applied to non-linear models.

The accuracy improvement also was particularly high when temporarily changing attributes were missing and when heterogeneity of data generation process was present.

The research in progress includes the performance analysis (convergence time, stability, etc.) of the proposed method and a comparison of the proposed and other known spatial-temporal regression methods on real-life data of various types.

Bibliography

[1] G. Box,G. Jenkins, and G. Reinsel, *Time Series Analysis, Forecasting and Control*,Third ed., Prentice Hall,Upper Saddle Hill,NJ,1994.

[2] P. Burridge, *Testing for a common factor in a spatial auto-regression model*, Environment and Planning A, 13(1981), pp. 795-800.

[3] N. Cressie, *Spatial Statistics*,Willey, New York,NY,1993.

[4] N. Cressie and J. J. Majure, *Spatio-temporal statistical modeling of livestock waste in streams*, Journal of Agricultural, Biological, and Environmental Statistics, 2(1997), pp. 24-47.

[5] R. Davidson R. and J. Mc Kinon, *Estimation and Inference in Econometrics*,Oxford Univ.Press.,New York,NY,1993.

[6] J. L. Devore, *Probability and Statistics for Engineering and the Sciences*, Fourth ed., Int'l Thomson Publishing Company,Belmont,CA,1985.

[7] R. Florax and H. Folmer, *Specification and estimation of spatial linear regression models- Monte Carlo evaluation of pre-test parameters*, Regional Science and Urban Economics, 22(1992), pp. 405-432.

[8] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Second ed., Prentice Hall,Upper Saddle Hill,NJ,1999.

[9] A. Householder, *The Theory of Matrices in Numerical Analysis*, Dover,New York,NY,1975.

[10] R. Pace Kelly and O. Gilley, *Generalizing the OLS and grid estimators*, Real Estate Economics, 26(1998), pp. 331-347.

[11] R. Pace Kelly,R. Barry,J. Clapp, and M. Rodriquez, *Spatiotemporal auto-regressive models of neighborhood effects*, Journal of Real Estate Economics, 17(1998), pp. 15-33.

[12] D. Pokrajac,T. Fiez, and Z. Obradovic, *A Tool for controlled knowledge discovery in spatial domains*, in Proc. 14th European Simulation Multiconference (ESM) 2000, to appear.

[13] D. Pokrajac,Z. Obradovic, and T. Fiez, *Understanding the influence of noise, sampling density and data distribution on spatial prediction quality through the use of simulated data*, in Proc. 14th European Simulation Multiconference (ESM) 2000, to appear.

[14] G. Upton and B. Fingleton, *Spatial Data Analysis by Example, Vol. 1: Point Pattern and Quantitative Data*,Wiley, New York,NY,1985.

[15] S. Vucetic,T. Fiez, and Z. Obradovic, *Examination of the influence of data aggregation and sampling density on spatial estimation*, Water Resources Research (2000), in press.

[16] K. Wikle and N. Cressie, *A dimension-reduced approach to space-time Kalman filtering*, Biometrika, 86(1999), pp. 815-829.