Spatio-temporal Statistical Models

Spatial Statistics and Spatial Econometrics [ECO 324/524], Winter 2025

Anshuman Bunga (2021016)

Econometrics Lab, Department of Social Sciences and Humanities
Indraprastha Institute of Information Technology, Delhi

May 25, 2025

Overview

1. **Know Thy Data**

2. **Spatio-temporal Prediction**

3. **Regression (Trend-Surface) Estimation**

4. **Model Diagnostics: Dependent Errors**

0. Why Bother With Spatio-temporal Methods?

- The rise of big-data has made spatio-temporal datasets (data indexed in both space and time) and tools readily available for analysis.
- Time series analysis models temporally-varying phenomenon. Spatial analysis models spatially-varying phenomenon. *What if the topic of interest is dependent on both space and time?*
    - Time series models explain patterns and behavior over time, while spatial models explain patterns across space. When focusing on time or space alone, *the influence of the dimension left out often remains unaccounted for, appearing as unexplained variance in the model's standard errors and residuals!*
    - Weaker explanation and prediction power for phenomenon of interest (poor standard errors, unexplained dependence of data values)
- In the pursuit of better performing models, can use regressors that vary with space/time/both, i.e more freedom in selecting model regressors.[1]

[1] Readers are encouraged to look up parameter inference, a statistical exercise that reveals whether some regressors are actually important in the model for explanation.

1. Know Thy Data

Introduction to the Groundwater Level Dataset

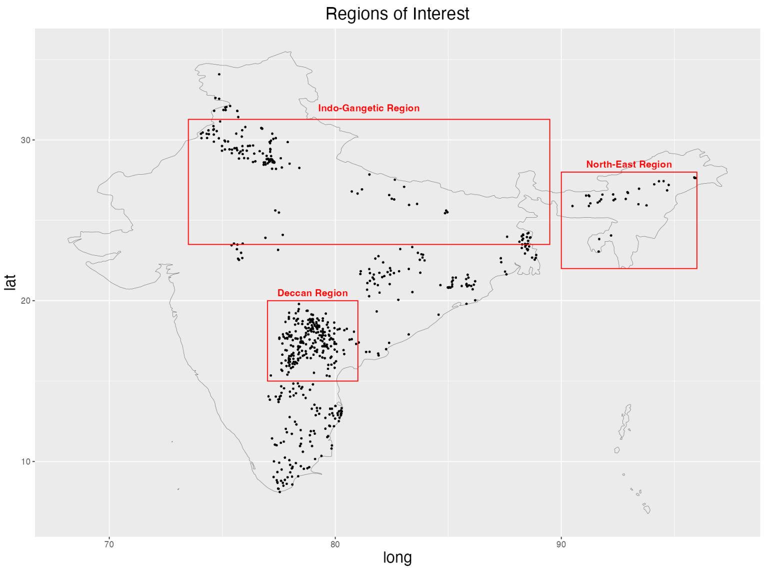
- Sourced from *India Water Resource Information Systems* (India-WRIS), created by the Ministry of Water Resources, India.
- Contains the ground water level measured at 18,849 measuring stations situated all over India over the period May 2023-May 2024.
- Came with 4,308,184 observations, with a host of its own issues (irregular observation frequency across stations, multiple observations at the same spatio-temporal location, nonsense observations- *what does discovering groundwater at a depth of -3583.09m mean? How about +33194.93m?* [2])

[2] As of 2020, the peak of Mount Everest is at the height +8848m.

Filtering and Balanced Panels

- We first applied a filter, restricting ourselves to observations that fell in the range (-10, -0.01)m.
    - Removing positive values got rid of around 300k observations. Applying the lower bound of -10 led to *halving* the dataset (left with 2,082,664 observations!)
    - A small allusion to how significant the exploratory data analysis step can be.
- We then found the stations that made observations in every month in the period (May 2023-Apr 2024), averaged their observations in each month, and successfully prepared a balanced panel.
    - 645 stations followed the required pattern.
    - Final size of balanced panel = 645 * 12 = **7740** observations.

Region of Interest



2. Spatio-temporal Prediction

Spatio-Temporal Prediction Problem

- To start with, consider the prediction (i.e., “interpolation”) of groundwater level in October 2023 at Delhi.
- Prediction approaches:
  - We somehow just combine the nearest observations in space, a.k.a. *Tobler’s Law*
  - We should consider nearby observations in both space *and* time.
- Types of spatio-temporal predictors:
  - Smoothing predictor: If we have data before and after October 2023 Our problem requires this type of predictor!
  - Filtering predictor: If we only have data before October 2023
  - Forecasting predictor: If we have to predict at any location, at any time after which data is not available

Deterministic Prediction: Inverse Distance Weighting

- Perhaps the simplest approach: follow Tobler's law and give more weight to the nearest observations in space and time
- Inverse distance weighting (IDW) for spatio-temporal data:

$$\{Z(\mathbf{s}_{11};t_1), Z(\mathbf{s}_{21};t_1), \ldots, Z(\mathbf{s}_{m_1 1};t_1), \ldots, Z(\mathbf{s}_{1T};t_T), \ldots, Z(\mathbf{s}_{m_T T};t_T)\}$$

  where for each time t_j, we have m_j observations (12 timestamps, 645 locations)
- IDW predictor at location $\mathbf{s}_0$ and time t_0 (assuming $t_1 \leq t_0 \leq t_T$):

$$\hat{Z}(\mathbf{s}_0;t_0) = \sum_{j=1}^{T} \sum_{i=1}^{m_j} w_{ij}(\mathbf{s}_0;t_0) Z(\mathbf{s}_{ij};t_j)$$

- Weights normalized to sum to 1:

$$w_{ij}(\mathbf{s}_0;t_0) \equiv \frac{\tilde{w}_{ij}(\mathbf{s}_0;t_0)}{\sum_{k=1}^{T} \sum_{\ell=1}^{m_k} \tilde{w}_{\ell k}(\mathbf{s}_0;t_0)}$$

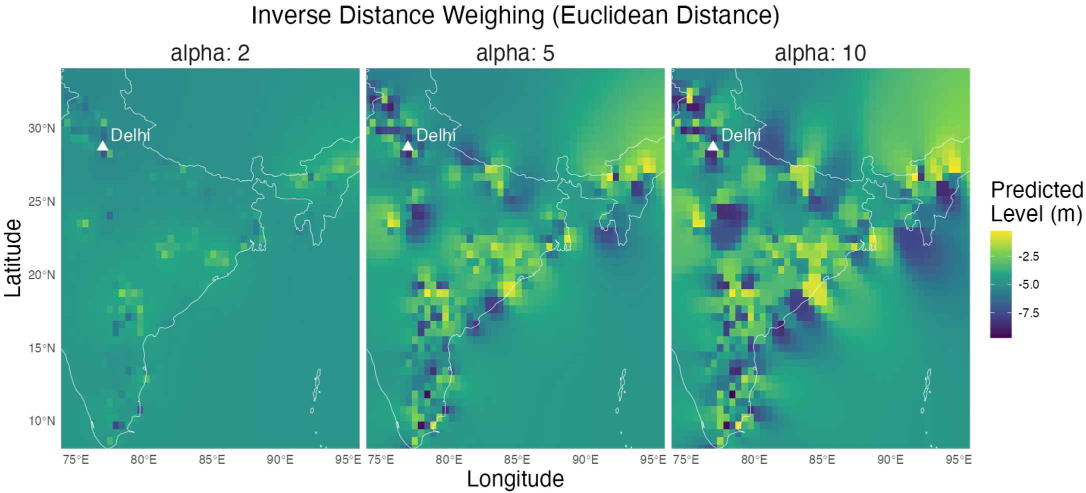
Inverse Distance Weighting (Continued)

- The unnormalized weights are proportional to the inverse of distance raised to power α:

$$\tilde{w}_{ij}(\mathbf{s}_0; t_0) \equiv \frac{1}{d((\mathbf{s}_{ij}; t_j), (\mathbf{s}_0; t_0))^{\alpha}}$$

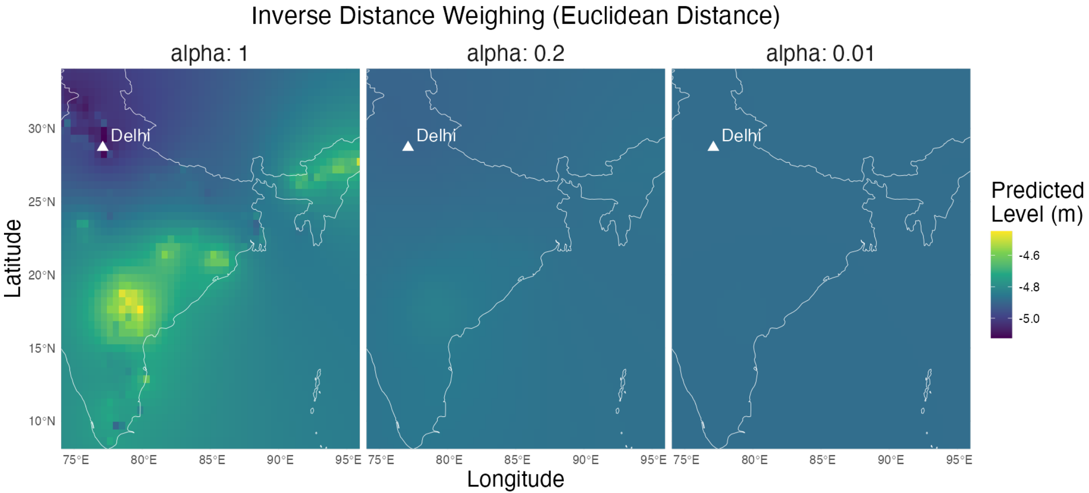
- Where:
  - $d((\mathbf{s}_{ij}; t_j), (\mathbf{s}_0; t_0))$ is the "distance" between the spatio-temporal locations
  - The power coefficient $\alpha > 0$ controls the amount of smoothing (often $\alpha = 2$)
- Key properties:
  - IDW is a weighted average of data points, giving closest locations more weight (taking the inverse of the distance encodes this idea in a mathematical manner)
  - Formula gives an exact interpolator: If prediction location matches a data location, prediction equals the data value
  - For a smoothing predictor: Use weights proportional to $\frac{1}{(d(\cdot,\cdot)+c)^{\alpha}}$ where $c > 0$

IDW Results (Euclidean Distance); $\alpha > 1$



- As α increases, distance increases, inverse distance decreases, and so does $\tilde{w}_{ij}$ accordingly. Nearer observations get a higher weight.

IDW Results (Euclidean Distance); $0 < \alpha \leq 1$



- As α decreases, inverse distance increases, and so does $\tilde{w}_{ij}$. The algorithm gives an almost equal weight to all observations for $\alpha \to 0$. *Note the lack of range in predicted values!*

Kernel-Based Prediction

- IDW is a type of spatio-temporal kernel predictor
- General formulation of unnormalised weights using kernel function:

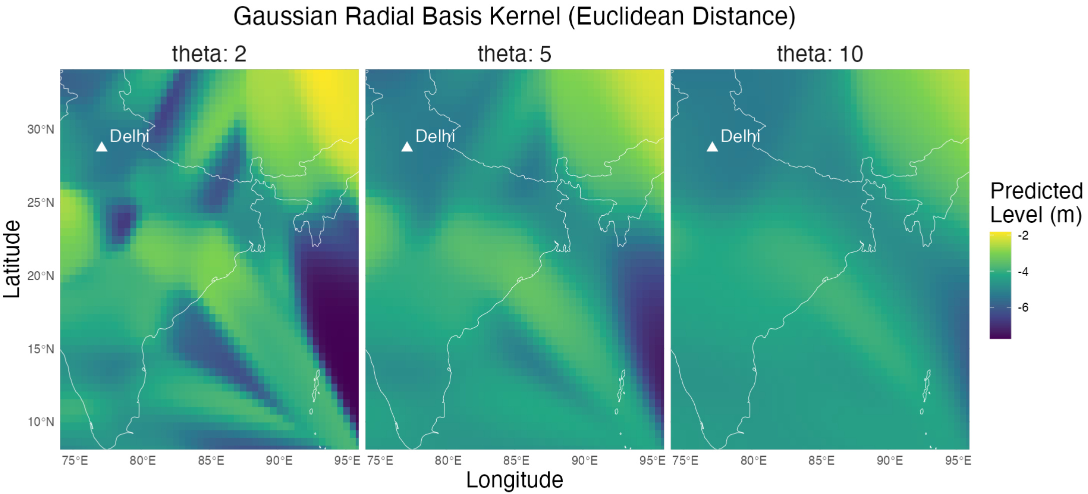
$$\tilde{w}_{ij}(\mathbf{s}_0; t_0) = k((\mathbf{s}_{ij}; t_j), (\mathbf{s}_0; t_0); \theta)$$

- Kernel function quantifies similarity between two locations:
  - Depends on distance between $(\mathbf{s}_{ij}; t_j)$ and $(\mathbf{s}_0; t_0)$
  - Controlled by bandwidth parameter θ
  - Larger bandwidth averages more observations (smoother predictions)
- Example: Gaussian radial basis kernel

$$k((\mathbf{s}_{ij}; t_j), (\mathbf{s}_0; t_0); \theta) \equiv \exp\left(-\frac{1}{\theta} d((\mathbf{s}_{ij}; t_j), (\mathbf{s}_0; t_0))^2\right)$$

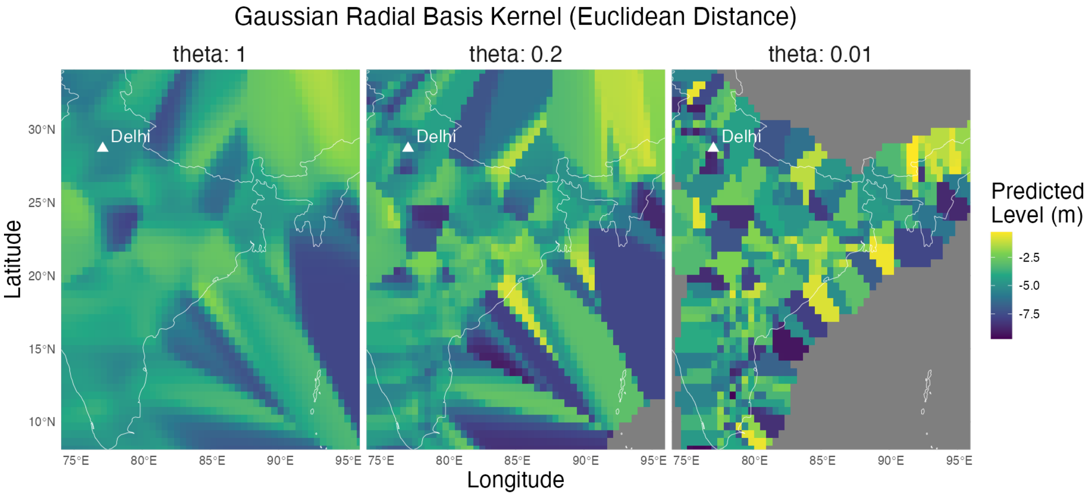
$$\theta \in (0, \infty)$$

Gaussian Kernel Prediction (Euclidean Distance); $\theta > 1$



- Note how the predictions become 'smoother' as the value of θ increases: higher values increase the "neighborhood" of points that contribute significantly to each prediction.

Gaussian Kernel Prediction (Euclidean Distance); $0 < \theta \leq 1$



- The model *isn't making predictions in the gray areas!* The kernel becomes extremely localized - it only considers very nearby points when making predictions. If the prediction points are too far from any observation in the dataset, the weights become essentially zero, resulting in no prediction.

Candidate Distance Functions (other than Euclidean)

- Weighted Combination of Spatial and Temporal Distance

  $$d(\mathbf{p}, \mathbf{q}) = w_s \cdot d_{\text{spatial}}(\mathbf{p}, \mathbf{q}) + w_t \cdot d_{\text{temporal}}(\mathbf{p}, \mathbf{q})$$

  where w_s and w_t are weighting parameters for spatial and temporal components respectively and $w_s + w_t = 1$
- Space-time Cube Distance

  $$d_{\text{ST}}(\mathbf{p}, \mathbf{q}) = \sqrt{\left(\frac{x_p - x_q}{s_x}\right)^2 + \left(\frac{y_p - y_q}{s_y}\right)^2 + \left(\frac{t_p - t_q}{s_t}\right)^2}$$

  where s_x, s_y, and s_t are scaling factors for each dimension, s_x and s_y normalize spatial units (e.g., meters, kilometers), s_t normalizes temporal units (e.g., seconds, days) to make them comparable with spatial units

Limitations of Deterministic Methods

- Do not explicitly account for measurement uncertainty
  - Non-exact interpolators do implicitly smooth observation errors, but there is *no mechanism to incorporate explicit knowledge of measurement error.*
- Do not provide model-based estimates of prediction uncertainty. The IDW algorithm gave you a predicted value for a spatio-temporal location. How good is the prediction, really? What's the confidence interval?
- Use cross-validation to find optimal parameters to IDW algorithms.
  - Use leave-one-out cross-validation (LOOCV) MSPE score
  - Different spatio-temporal kernel functions can potentially **outperform IDW** for a given dataset

3. Regression (Trend-Surface) Estimation

Regression (Trend-Surface) Estimation

- Alternative to deterministic predictors: use a basic statistical regression model
- Key advantages:
  - *Exceptionally simple to implement* in almost any software package
  - *Explicitly accounts for model error* (usually assumed independent)
  - Provides a *model-based prediction-error variance*
- Consider observations at discrete times $\{t_j : j = 1, \ldots, T\}$ for all spatial locations $\{\mathbf{s}_i : i = 1, \ldots, m\}$ Reminder: $T = 12$, $m = 645$!
- Basic model form:

$$Z(\mathbf{s}_i; t_j) = \beta_0 + \beta_1 X_1(\mathbf{s}_i; t_j) + \ldots + \beta_p X_p(\mathbf{s}_i; t_j) + \epsilon(\mathbf{s}_i; t_j)$$

where $\epsilon(\mathbf{s}_i; t_j) \sim$ indep. $N(0, \sigma_\epsilon^2)$

Covariates in Trend-Surface Models

- Covariates $X_k(\mathbf{s}_i; t_j)$ can represent various types of effects:
  - Spatially varying, temporally invariant features
    - Example: elevation, soil type
  - Time trends that are spatially invariant
    - Example: festivals, agricultural seasons
  - Spatially and temporally varying variables
    - Example: humidity, precipitation
- Can also incorporate spatio-temporal "basis functions" to reconstruct the observed data
- Assumes spatio-temporal dependence can be accounted for by covariate terms

Fitting Regression Models via OLS

- The regression model can be fitted via ordinary least squares (OLS)
- OLS estimates parameters $\beta_0, \beta_1, \ldots, \beta_p$ by minimizing the residual sum of squares:

$$RSS = \sum_{j=1}^{T} \sum_{i=1}^{m} (Z(\mathbf{s}_i; t_j) - \hat{Z}(\mathbf{s}_i; t_j))^2$$

- Estimated regression equation:

$$\hat{Z}(\mathbf{s}; t) = \hat{\beta}_0 + \hat{\beta}_1 X_1(\mathbf{s}; t) + \ldots + \hat{\beta}_p X_p(\mathbf{s}; t)$$

- Also obtain an estimate of error variance: $\hat{\sigma}_e^2 = \frac{RSS}{(mT-p-1)}$
- Benefits:
  - Predictions for mean response at any location with covariates
  - Model-based uncertainty estimates for predictions
- However, our regression equation does not explicitly account for measurement errors in the responses, and thus the variation due to measure error is *confounded with the variation due to lack of fit in the residual variance* σ_e^2.

Choosing an OLS Model

- OLS is always the starting point for regression tasks due to its desirable statistical features under the OLS model assumptions (**BLUE**- best linear unbiased estimator)
- Spatio-temporal data highly likely to violate the assumptions of the OLS model, in which case we then look for more sophisticated estimators.

Choosing an OLS Model

- We will perform an OLS regression using the following equation using R:

$$\begin{aligned}\text{groundwater}(s_i;t_j) = \beta_0 &+ \beta_1\text{latitude}(s_i;t_j) + \beta_2\text{longitude}(s_i;t_j) + \beta_3\text{time}(s_i;t_j)\\ &+ \beta_4\text{lat\_time}(s_i;t_j) + \beta_5\text{long\_time}(s_i;t_j) + \beta_6\text{lat\_long}(s_i;t_j)\\ &+ \beta_7 D_{\text{monsoon}}(s_i;t_j) + \beta_8 D_{\text{indo\_gangetic}}(s_i;t_j)\\ &+ \beta_9 D_{\text{north\_east}}(s_i;t_j) + \beta_{10} D_{\text{deccan}}(s_i;t_j) + \epsilon(s_i;t_j)\end{aligned}$$

- Where:
    - *lat_time* denotes the latitude-time interaction term
    - *long_time* denotes the longitude-time interaction term
    - *lat_long* denotes the latitude-longitude interaction term
    - $D_{\text{monsoon}} = 1$ if the observation time falls in the range June-September, 0 otherwise.
    - $D_{\text{indo\_gangetic}} = 1$ if the location falls in the Indo-Gangetic region, 0 otherwise
    - $D_{\text{north\_east}} = 1$ if the location falls in North-East India, 0 otherwise
    - $D_{\text{deccan}} = 1$ if the location falls in the Deccan plateau, 0 otherwise

OLS Regression Results

OLS Model Beta Values, Standard Errors and Significance Level

	DV: reading
y	0.380^{*}
	(0.224)
x	0.082
	(0.071)
time	-0.456^{***}
	(0.142)
lat_time	0.001
	(0.001)
long_time	0.007^{***}
	(0.002)
lat_long	-0.006^{*}
	(0.003)

	DV: reading
monsoon	0.029
	(0.056)
indo_gangetic	0.347^{***}
	(0.110)
north_east	1.830^{***}
	(0.236)
deccan	0.360^{***}
	(0.065)
Constant	-10.890^{**}
	(5.547)

${}^{*}p < 0.1$; ${}^{**}p < 0.05$; ${}^{***}p < 0.01$

Inference on OLS Model Parameters

Spatial Covariates:

- Significant **regional variations**: North-East region shows highest effect (1.830^{***}). Indo-Gangetic plain (0.347^{***}) and Deccan plateau (0.360^{***}) show similar magnitudes: readings in these areas are expected to be higher.
- **Latitude-longitude interaction** slightly negative (-0.006^{*}): effect of longitude becomes more negative as latitude increases, is moderately significant

Temporal Covariates:

- **Time** has significant negative main effect (-0.456^{***}), showing that readings are decreasing over time.
- **Monsoon** has a small positive coefficient, suggesting slightly higher readings during monsoon periods. However, this effect is not statistically significant.

Spatio-temporal Covariates

- **Longitude-time interaction** positive (0.007^{***}), suggests the effect of longitude becomes more positive over time. Highly significant interaction.

Inference on OLS Model Statistics

Table: OLS Regression Model Statistics

Observations	7,740
R^2	0.043
Adjusted R^2	0.042
Residual Std. Error	2.279 (df = 7729)
F Statistic	35.071^{***} (df = 10; 7729)
Note:	*p<0.1; **p<0.05; ***p<0.01

- Highly significant F-statistic indicates that the model has explanatory power.
- However, a poor R^2 statistic tells us that the model explains only 4.3% of the total variance in groundwater level - limited predictive power!

Alternate OLS Models

- According to intuition, $D_{monsoon}$ may have had a significant positive impact on the groundwater level in an area. An alternate OLS model uses this dummy variable as the *only* temporal covariate:

$$\begin{aligned}\text{groundwater}(s_i; t_j) = \beta_0 &+ \beta_1 \text{latitude}(s_i; t_j) + \beta_2 \text{longitude}(s_i; t_j) \\ &+ \beta_3 \text{lat\_time}(s_i; t_j) + \beta_4 \text{long\_time}(s_i; t_j) + \beta_5 \text{lat\_long}(s_i; t_j) \\ &+ \beta_6 D_{\text{monsoon}}(s_i; t_j) + \beta_7 D_{\text{indo\_gangetic}}(s_i; t_j) \\ &+ \beta_8 D_{\text{north\_east}}(s_i; t_j) + \beta_9 D_{\text{deccan}}(s_i; t_j) + \epsilon(s_i; t_j)\end{aligned}$$

Alternate OLS Regression Results

OLS Model Beta Values, Standard Errors and Significance Level

	DV: reading
y	0.383*
	(0.224)
x	0.118*
	(0.070)
lat_time	0.0002
	(0.001)
long_time	0.001***
	(0.0003)
lat_long	−0.006*
	(0.003)
monsoon	0.027
	(0.056)
indo_gangetic	0.347***
	(0.110)
north_east	1.830***
	(0.236)
deccan	0.360***
	(0.065)
Constant	−13.852**
	(5.473)

${}^{*}p < 0.1$; ${}^{**}p < 0.05$; ${}^{***}p < 0.01$

- The $D_{monsoon}$ variable remains statistically insignificant! Longitude variable x becomes moderately significant in the new model, other inferences remain intact.

Inference on Alternate OLS Model Statistics

Table: OLS Regression Model Statistics

Observations	7,740
R^2	0.042
Adjusted R^2	0.041
Residual Std. Error	2.281 (df = 7730)
F Statistic	37.770*** (df = 9; 7730)
Note:	*p$<$0.1; **p$<$0.05; ***p$<$0.01

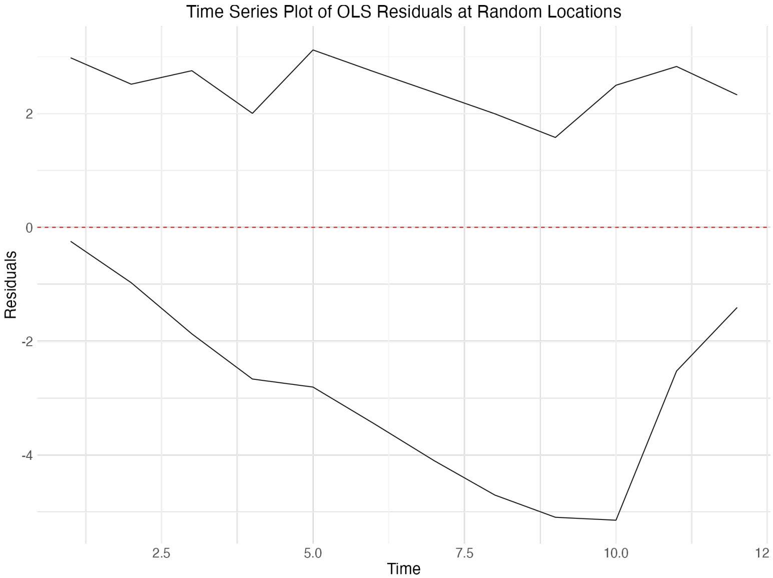
- Highly significant F-statistic remains, also higher in magnitude compared to previous model.
- R^2 and residual standard errors deteriorate in this model!

4. Model Diagnostics: Dependent Errors

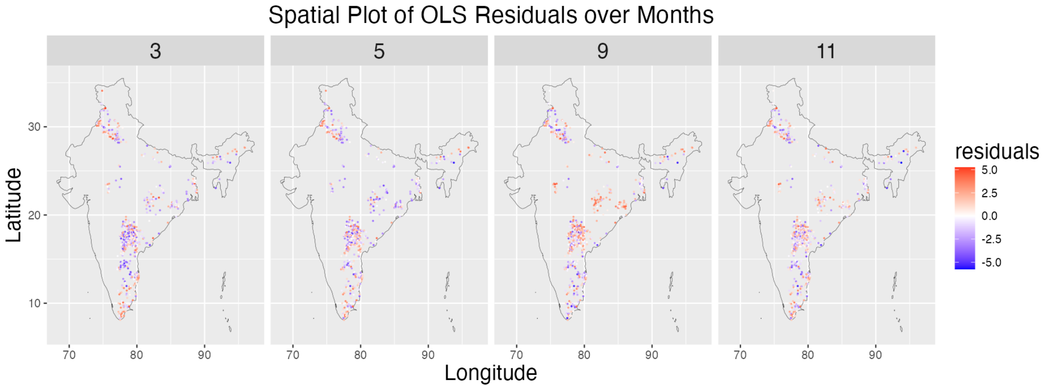
Model Diagnostics: Dependent Errors

- When we first learn how to do regression modeling in statistics, we gain an *appreciation for the importance of model diagnostics* to verify the assumptions of the model
- Standard regression diagnostics examine:
  - Presence of outliers
  - Influential observations
  - Non-constant error variance
  - Non-normality of errors
  - Dependence in the errors
- It is *particularly important to consider the possibility of dependent errors* in the case where data is indexed in space or time, or both. *Our original motivation behind spatio-temporal models!*

OLS Residuals - Plot over Time



OLS Residuals - Plot over Space



Reading the Residual Plots

- Plot over time:
  - The two locations were chosen at random from the set of locations. For one location the model underestimates groundwater level, for the other location groundwater level is overestimated[3].
  - Note the trend present in the residuals.
- Plot over space:
  - One can visualise *some extent* of spatial dependence in the plots. We can try to use more sophisticated tools to confirm spatial dependence, though.

[3] We take some time here to consider the sign of the residuals: a positive residual means that groundwater was predicted to be located *deeper* than the observed height, and a negative residual means that groundwater was predicted to be located at a height *above* the observed value!

Intrinsic Stationarity: The Variogram

- Consider for a moment that we are conducting a *spatial* analysis of a random process $\{Z(s) : s \in D\}$ where D is the index set of locations s, called the domain.
- If we **decide**[4]that the random process is intrinsically stationary, then the following mathematical statements hold:

$$E(Z(s+h) - Z(s)) = 0 \tag{1}$$

$$var(Z(s+h) - Z(s)) = 2\gamma(h) \tag{2}$$

- The quantity $2\gamma(h)$ is called the spatial variogram, defined for a spatial lag h, a separation vector. The quantity $\gamma(h)$ is called the semivariogram.
- Intrinsic stationarity comes with the assumptions that the population mean is spatially invariant (1) and that the variance of the difference between any two observations depends only on the spatial lag h (2).

[4] Stationarity is a *decision* taken by a statistician about the random process, not a hypothesis that can be tested!

Empirical Spatial Variogram

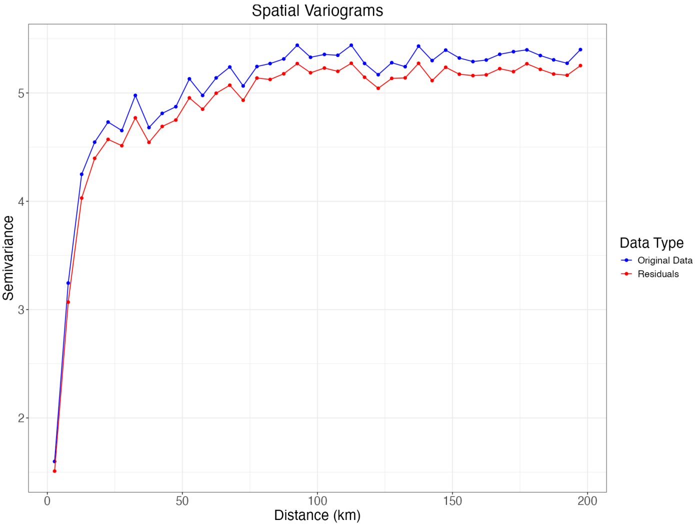
- Let us consider that our groundwater data to be temporally invariant. We can calculate the empirical variogram for our data using the following estimator proposed by *Matheron (1962)*:

$$2\hat{\gamma}(h) = \frac{1}{|N(h)|} \sum_{N(h)} (Z(s_i) - Z(s_j))^2$$

$$N(h) \equiv \{(i,j) : s_i - s_j = h\}$$

- A variogram is fundamentally telling you how similar or different locations are as a function of how far apart they are.
- Low variogram values imply that locations are similar (highly correlated), while high variogram values imply that locations are dissimilar (weakly correlated).

Empirical Spatial Semivariogram (Groundwater Data)



Temporal Variograms

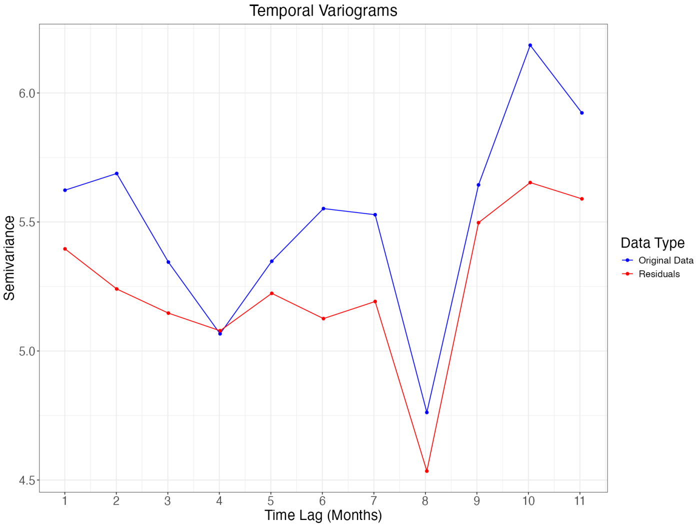
- Time series analysis generally uses tools like the autocovariance function and the autocorrelation function.
- Calculating the temporal variogram can also help us see how values at a fixed location vary over increments in time (instead of increments in distance.)

$$2\gamma(\tau) = var(Z(t+\tau) - Z(t))$$

$$2\hat{\gamma}(\tau) = \frac{1}{|N(\tau)|} \sum_{N(\tau)} (Z(t_i) - Z(t_i + \tau))^2$$

$N(\tau)$ is the number of pairs of observations separated by a temporal lag τ.

Empirical Temporal Semivariogram (Groundwater Data)



Reading Variogram Plots

- If a model captures the spatial/temporal/spatiotemporal patterns present in the data, one can expect the variogram of the data to differ from the variogram plotted from the model's resiudals.
- The spatial semivariograms show that the semivariance stabilises at a distance, however both the data and the residual semivariogram show the same pattern.
  - Possible spatial structure of data might not have been fully captured by our OLS model!
- Recall that the definition for a temporal variogram includes a temporal lag τ. The plot will help us answer questions about the similarity between values that are τ months apart.
  - Our plot suggests high correlation of groundwater levels between values with a time lag of 8 months. For example, values in May 2023 and May 2023 + 8 months (i.e. Jan 2024) are highly correlated. The lowest correlation can be observed with values having a time lag $\tau = \{2, 10\}$.
  - The residual variogram is flatter than the original data's variogram, however both show the same seasonal trends. This tells us that our simple model did not pick up on the seasonal trends present in the data!

The Spatio-temporal Covariogram

- Our detour into spatial and temporal variograms, as well the conditions of stationarity allows us to now construct an exploratory tool to analyse for dependence called the spatio-temporal variogram, given as

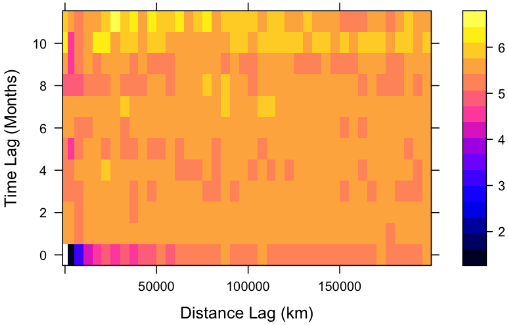
$$2\gamma_z(\mathbf{h};\tau) = \mathrm{var}(Z(\mathbf{s}+\mathbf{h};t+\tau) - Z(\mathbf{s};t))$$

  where **h** is a spatial lag and τ is a temporal lag.
- From an exploratory perspective, we can calculate the empirical spatio-temporal covariogram from both the original data and the residuals for the purpose of comparison:

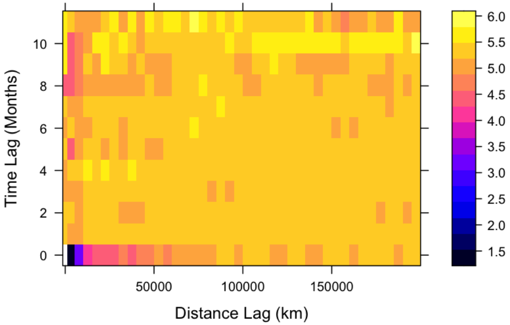
$$\hat{\gamma}_z(\mathbf{h};\tau) = \frac{1}{|N_s(\mathbf{h})||N_t(\tau)|} \sum_{\mathbf{s}_i,\mathbf{s}_k \in N_s(\mathbf{h})} \sum_{t_j,t_\ell \in N_t(\tau)} (Z(\mathbf{s}_i;t_j) - Z(\mathbf{s}_k;t_\ell))^2$$

Empirical Spatio-Temporal Semivariogram Comparison

Spatio-temporal Variogram (Original Data)



Spatio-temporal Variogram (Residuals)



The residuals have smaller semivariance values but about the same visualisation as the original data, which tells us that the model did not capture the spatio-temporal correlation patterns present in the data.

Statistical Tests for Dependence

- Visualisation and exploratory techniques are a great starting point for identifying patterns and trends. One could claim dependence in the residuals beyond a *reasonable* doubt. A picture is worth a thousand words.
- However, we still require some mathematical tools to prove the existence of dependence beyond *all* doubt. For this purpose, we employ statistical tests.
- Tests for temporal dependence:
  - Durbin-Watson test
- Tests for spatial dependence:
  - Moran's I test for areal regions
- Tests for spatio-temporal dependence:
  - Spatio-temporal Analog to Durbin-Watson
  - "Space-time index" (STI) approach (Henebry, 1995)
  - Extension of Moran's I statistic for spatio-temporal data

Durbin-Watson Test for Autocorrelation

- The Durbin-Watson (DW) test detects *temporal autocorrelation* in regression residuals:

  $$DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}, DW \in (0, 4)$$

  Values < 2 suggest positive autocorrelation (common in time series), while values > 2 suggest negative autocorrelation. Values close to 2 suggest no autocorrelation.
- Statistical inference:
  - Null hypothesis (H_0): No autocorrelation in residuals
  - Alternative hypothesis (H_A): Residuals exhibit autocorrelation
- Bonferroni correction for multiple tests:
  - Adjusted significance level: $\alpha_{corrected} = \alpha/m$ (for m tests)
  - **For our groundwater level data**: With 645 spatial locations, use $\alpha = 0.8612/645 = 0.0013$, implying strong positive autocorrelation.
  - **Drawback**: *Conservative* for spatial data due to possible presence of spatial autocorrelation

Moran's I Test for Spatial Autocorrelation

- Moran's I is a key statistic for measuring spatial[5] autocorrelation:

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, I \in (-1, 1)$$

  Values closer to -1 imply negative autocorrelation, values closer to $+1$ imply positive autocorrelation.

- Statistical inference:
  - Null hypothesis (H_0): No spatial autocorrelation (spatial randomness)
  - Z-scores and p-values determine significance of the observed pattern, with small p-values (e.g., $p < 0.05$) indicating significant spatial structure
  - **For our groundwater level data**: Global Moran's I $= 0.7666371$, max $p = 1.236e$-03 provides strong evidence to reject H_0

[5] For spatio-temporal autocorrelation, the distance between two spatio-temporal locations can be encoded using any candidate distance function (with both space and time parameters)

Limitations of Linear Models

- It is very common, when studying environmental phenomena, that a linear model with covariates will not explain all observed spatio-temporal variability
- Fitting such models frequently results in residuals that are spatially and temporally correlated
- Not surprising since environmental processes are more complex than can be described by simple geographical and temporal trend terms
- Evidence of dependence in residuals:
  - Temporal correlation: residuals close in time tend to be more similar than those far apart
  - Spatial correlation: residuals close in space tend to be more similar than those far apart
  - Statistical tests (Durbin-Watson, Moran's I) confirm these patterns

GLS and Associated Challenges

- Once diagnostics suggest spatio-temporal dependence in errors, what can we do?
- Generalized Least Squares (GLS) explicitly accounts for dependence in errors
    - Relaxes the assumption of independence in the errors
    - Allows $e(\mathbf{s}_i; t_j)$ and $e(\mathbf{s}_\ell; t_k)$ to be correlated
    - Error vector $\mathbf{e} \equiv (e(\mathbf{s}_1; t_1), \ldots, e(\mathbf{s}_m; t_T))'$ has distribution $\mathbf{e} \sim N(\mathbf{0}, \mathbf{C}_e)$ where $\mathbf{C}_e$ is a spatio-temporal covariance matrix
- Major challenge: Do we know the covariance matrix in advance?
    - Typically, no
- For prediction at locations without data:
    - Need to know error dependence between any two locations in time and space
    - Not just at observation points
    - Requires a model for the covariance structure
- Additional complications:
    - Computational challenges with large datasets
    - Need to estimate covariance parameters
    - Model misspecification risks

The End[6]

[6] of the presentation. Your next step lies in the references!

References

[1] Wikle, C.K., Zammit-Mangion, A., & Cressie, N. (2019). *Spatio-Temporal Statistics with R*. Chapman & Hall/CRC, Boca Raton, FL. URL: https://spacetimewithr.org/index

[2] Cressie, N. (1993) *Statistics for Spatial Data*. John Wiley & Sons, Inc., New York, USA URL: https://onlinelibrary.wiley.com/doi/book/10.1002/9781119115151

[3] R Core Team (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/ [7]

[7] The R code used for the models and plots can be found in the presentation's associated Github repository. https://github.com/AstraTriesGit/STSE-Groundwater