UNIVERSITÀ DEGLI STUDI DI BERGAMO
Dipartimento di Ingegneria Gestionale, dell'Informazione e della Produzione

Geo-Temporal Analysis on Salt Lake city traffic

Silviu Filote - 1059252
Jonathan Bommarito - 1068755

Statistics for High Dimensional Data (S4HDD) and CompStat Lab a.a. 2023/2024 (2nd edition)

DATA Febbraio 2024

Outline

1. Introduction
2. Dataset description and enrichment
3. Research Focus and Methodology Overview
4. DCM modeling
5. Seasonality management in DCM models
6. HDCM modeling
7. Conclusion
8. References

Introduction

Nowadays, traffic congestion represents one of the most common and persistent challenges afflicting urban environments worldwide. The main reason could be public transports and mobility in most parts of the world are not really developed, also the infrastructure and technology required in order to improve the transportation system seems to be really complex and expensive. Furthermore, the necessity of relying on private or personal vehicles for mobility amplify the traffic congestion. The analysis of this research aims to apply geo-temporal models and examine how they perform on the Salt Lake City traffic recorded by stations placed around the city, taking care of understanding which are the main factors that plague the traffic congestion.

Dataset description and enrichment

Dataset description

- Selected dataset is called “Salt Lake City Traffic” and it’s available on Kaggle
- The original dataset is divided into 2 parts
    - **"Utah Traffic":** Hourly traffic observations for 50 stations from January 1, 2022, 00:00, to January 31, 2022, 23:00. Missing data identified as NaN values.
    - **"Utah Traffic MetaData"**: Contains latitude, longitude information for each station, and the route name.
- Enriching the dataset due to lack of usable variables as covariates.

Dataset enrichment

- Initial dataset: Covariate limited to station route names
- Introduction of additional dummy variables:
    - "**Weekend**": Indicates Saturday or Sunday (1 if true, 0 otherwise).
    - "**Holidays**": Identifies significant holidays in the month.
    - "**Traffic on**": Reflects peak traffic hours (1 between 7 AM and 5 PM, 0 otherwise).
    - "**Hours**": Helps explain seasonality.
- Classification of stations based on route names and integration as dummy variables:
    - "**Interstate**": High-speed roads across multiple states.
    - "**US**": U.S. Route national highways connecting cities and regions.
    - "**RS**": State Route roads linking rural areas, cities, and locales within a state.
- Precipitation and Temperature Enrichment at all spatial locations via open-meteo.com API as temporal covariates, due to low spatial dependency:
    - "**mean prec**": mean of the precipitations
    - "**mean temp**": mean of the temperature

Research focus and methodology overview

Overview

- **Objective:** examine Salt Lake City traffic as a response variable using DCM and HDGM models along with enrichment the covariates
- **Model Performance Evaluation:**
  - residuals Analysis
  - validation phase using a clustering method and evalutation of the validation statistics
  - Log-likelihood Model's Parameter
  - kriging performance in estimating unknown areas near the network
- **Tuning Approaches:**
  - Maximization of log-likelihood function
  - EM convergence based on randomization of initial parameters
- All analyses conducted and model implementations are performed using D-STEM v2 Matlab library

DCM modeling

The DCM

- $\boldsymbol{x_\beta}(\boldsymbol{s}, t)$ and $\boldsymbol{x_z}(\boldsymbol{s})$ are vectors of covariates. Note that $\boldsymbol{x_z}(\boldsymbol{s})$ is time invariant
- $\omega(\boldsymbol{s}, t) \sim GP(0, \rho(||\boldsymbol{s} - \boldsymbol{s}'||; \boldsymbol{\theta}))$ is correlated over space but IID over time
- $\boldsymbol{z}(t)$ is $q \times 1$ dimensional with Markovian dynamics
- G is a stable $q \times q$ transition matrix
- $\boldsymbol{\eta}(t) \sim N(\boldsymbol{0}, \boldsymbol{\Sigma_\eta})$ is the innovation with $\boldsymbol{\Sigma_\eta}$ the variance-covariance matrix
- $\varepsilon(\boldsymbol{s}, t) \sim N(0, \sigma_\varepsilon^2)$ is the measurement error
- The model parameter set is $\Psi = \{\boldsymbol{\beta}, \alpha, \sigma_\varepsilon^2, \boldsymbol{\theta}, \boldsymbol{G}, \boldsymbol{\Sigma_\eta}\}$

$$y(\boldsymbol{s}, t) = \boldsymbol{x_\beta}(\boldsymbol{s}, t)'\boldsymbol{\beta} + \boldsymbol{x_z}(\boldsymbol{s})'\boldsymbol{z}(t) + \omega(\boldsymbol{s}, t) + \varepsilon(\boldsymbol{s}, t)$$

$$\boldsymbol{z}(t) = \boldsymbol{G}\boldsymbol{z}(t-1) + \boldsymbol{\eta}(t)$$

$$\omega(\boldsymbol{s}, t) = \sum_{j=1}^{c} \alpha_j x_j(\boldsymbol{s}, t)\omega_j(\boldsymbol{s}, t)$$

Formula: DCM model description

DCM tuning

- **Objective:** explore various parameterizations of the response variable using DCM for optimal performance selection.
- All the models analyzed have the same initial parametrization
- Based on the training and validation statistics the `DCM_m4` seems to be a good overall model

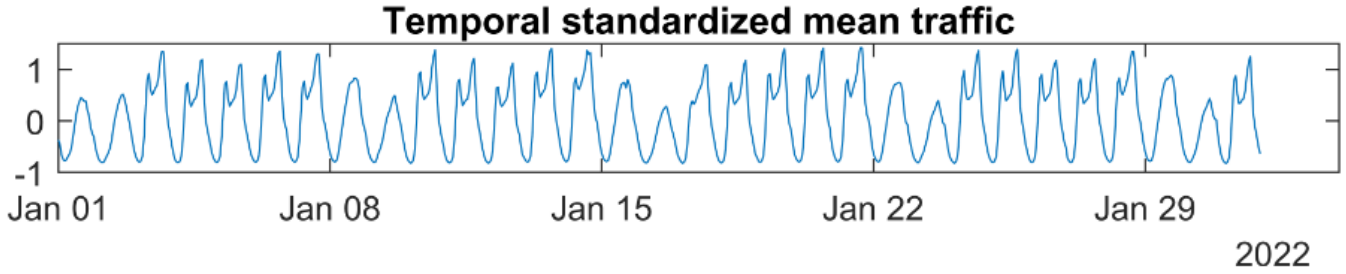


Figure 1: Temporal mean of the standardized traffic

Models	$\overline{R^2_{t,s}}$	$\overline{R^2_{v,s}}$	$\overline{R^2_{v,t}}$	$\overline{RMSE}_{v,t}$	$\overline{RMSE}_{v,s}$	$logL$	$lbqtest$	$archtest$
DCM_m1	-14.41	-3.46	0.33	0.44	0.39	4.96e+03	1	1
DCM_m2	0.78	-4.77	0.11	0.52	0.47	1.28e+04	1	1
DCM_m3	-18.22	-4.86	0.09	0.52	0.48	4.19e+03	1	1
DCM_m4	0.37	-2.76	0.44	0.39	0.34	1.10e+04	1	1

Table 1: Most performant models built using a DCM modelization with the same initial parametrization. For each model validation and training parameters are reported along with the log-likelihood parameter as well as the test conducted on the residuals.

DCM implementation

- Implementing the `DCM_m4`
- Splitting into training (80%) and validation (20%), using clustering method on stations
- All enrichment variables used as covariates, but will remain only the significant ones checking on the t-statistic to avoid overfitting
- The best model construction is summarized as:
  - $x_\beta(s,t)$ contains all the covariates
  - $x_z(s)$ matrix of ones.
  - $x_p(s)$ incorporates only `interstate`

Table 2: All covariates included in the fixed effect part of the `DCM_m4` model

Loading coefficient	Value	Std	$\|t\|$
weekend	-0.241	0.033	7.253
holidays	-0.105	0.060	1.755
mean temp	-0.026	0.017	1.485
mean prec	0.010	0.008	1.234
traffic on	0.466	0.017	27.118
hours	0.101	0.007	13.696
interstate	0.319	0.030	10.468
US	-0.733	0.026	28.096
RS	-0.680	0.026	26.348

Validation phase

- Highlighted spikes in the R^2 parameters indicate the model's incapacity to fully explain the total variance present in the data
- Poor estimation properties

- Validation statistics reflect seasonality
- Inability to accurately capture the high traffic phenomena during the week peak traffic times

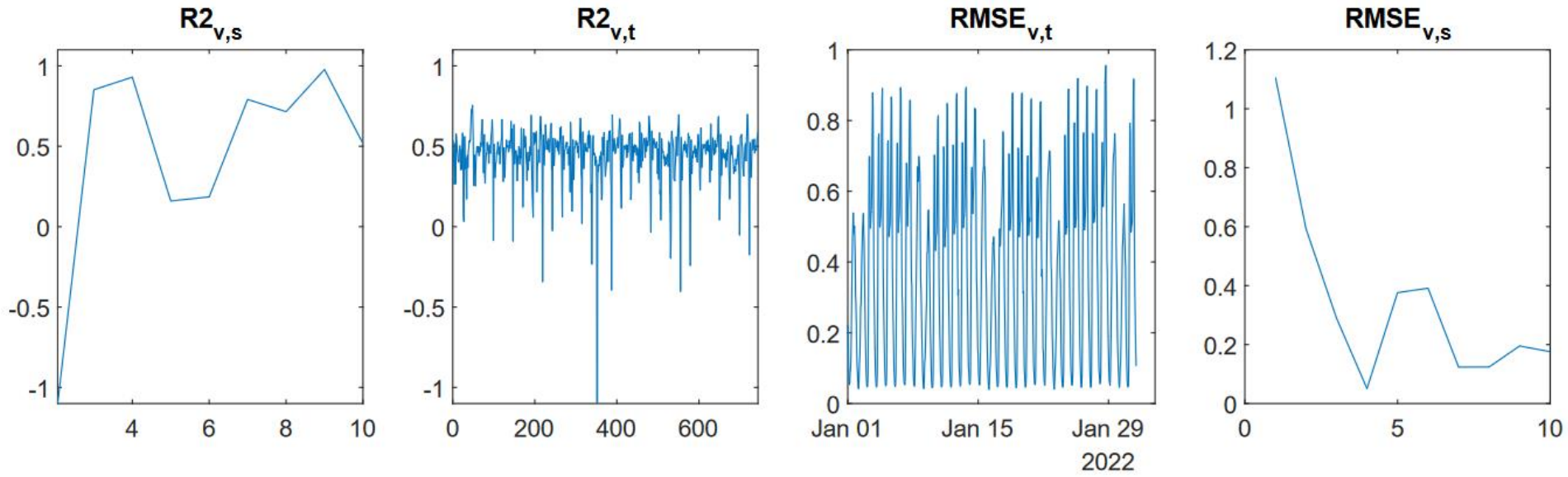


Figure 2: Temporal and spatial validation statics of the DCM_m4 model. The spikes defined as peaks surpassing the imposed y-axis limit

Latent variable

- The latent variable as temporal correction over the fixed effect
- Significant behaviour over time
- Seasonal pattern expected based on the traffic distribution

- Enrichment covariates as well as latent variable not enough to mitigate the seasonal component

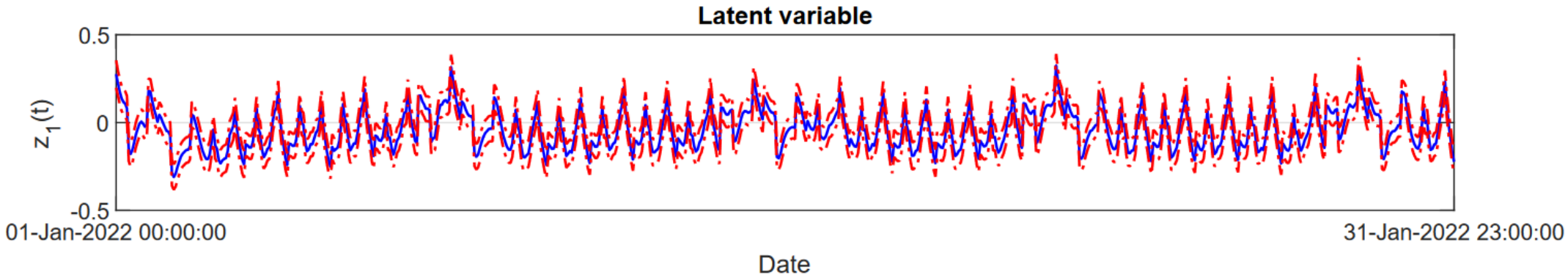


Figure 3: Latent variable estimated by the DCM_m4 model

Residual analysis

- `lbqtest` confirms correlation over the 24-hour period
- `Archtest` confirms eteroskedastic behavior
- Dubious residual distribution

- Evident seasonal pattern remaining within the residuals
- `DCM_m4` fails to capture all the variability present in the initial data, as evidenced by the residual attitude
- Model assumptions aren't met

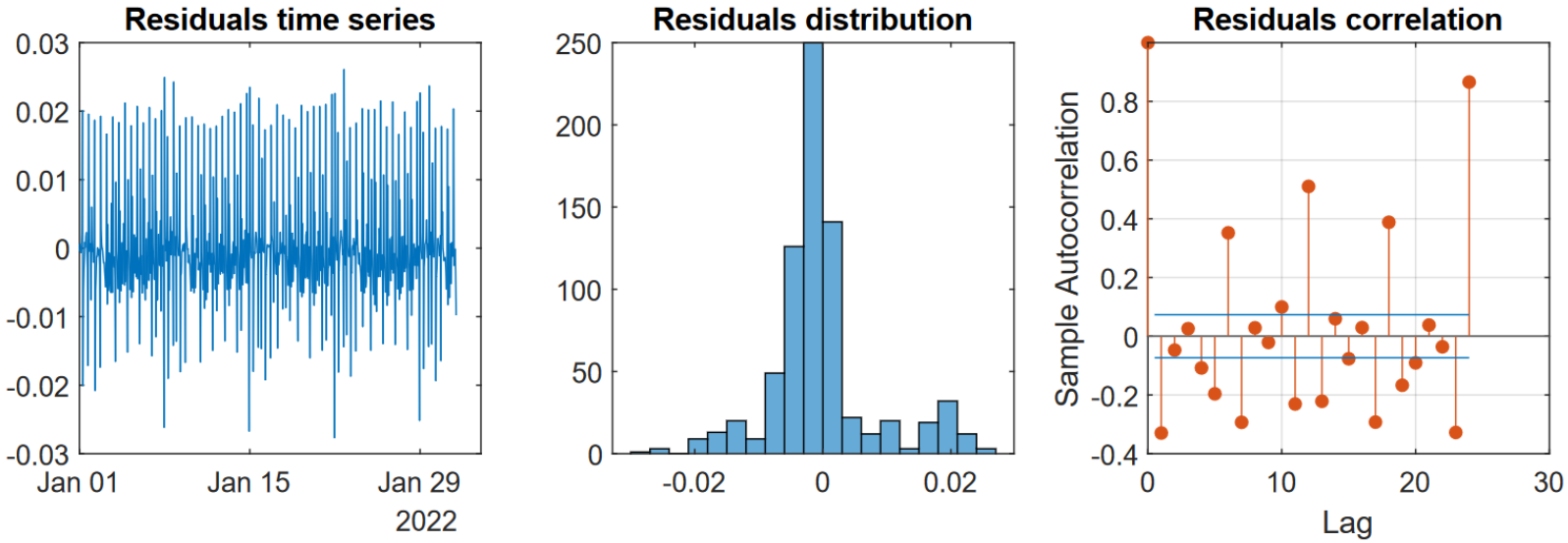


Figure 4: Temporal mean residual statistics derived from the training phase of the DCM_m4 model. During the training phase 80% of the available data is used, which includes 40 stations out of 50

Kriging results

- Evaluate the predictive capacity of the `DCM_m4` model
- Predict the traffic in 4 specific locations
- These locations coordinates were selected in proximity to our existing traffic network
- At these spatial locations choose by us we also added the enrichment covariates
- Really high kriging uncertainty compared to the mean traffic estimated

Table 3: Mean spatial estimates and uncertainty of the kriging locations. The locations are listed starting from the top to the bottom of the figure 6

Latitude	**Longitude**	$\overline{yhat}_k$	$\overline{\sigma}_k$
40.77	-112.14	35600.8	32238.3
40.54	-111.89	79909.4	21265.5
40.38	-111.96	11483.4	1499.44
40.10	-111.68	38140.5	1499.44

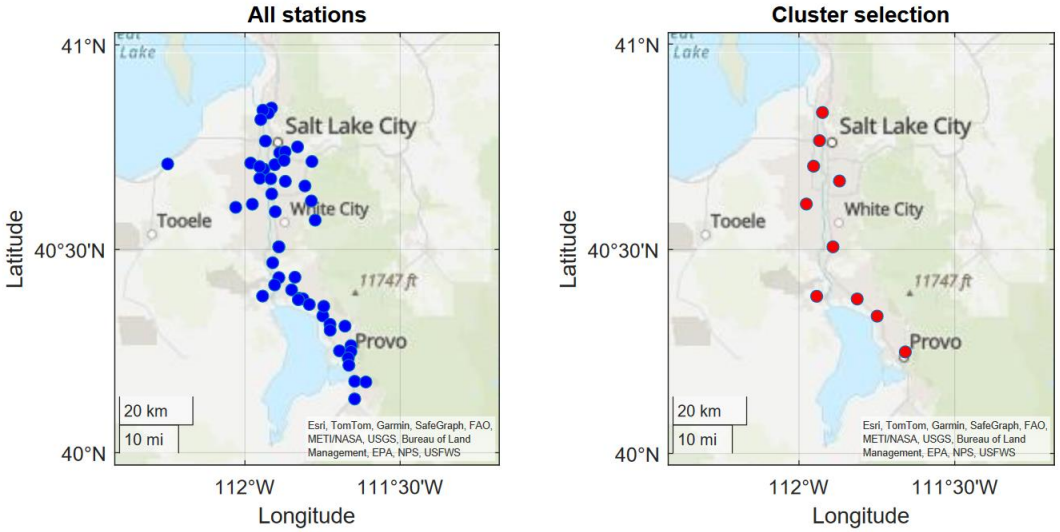
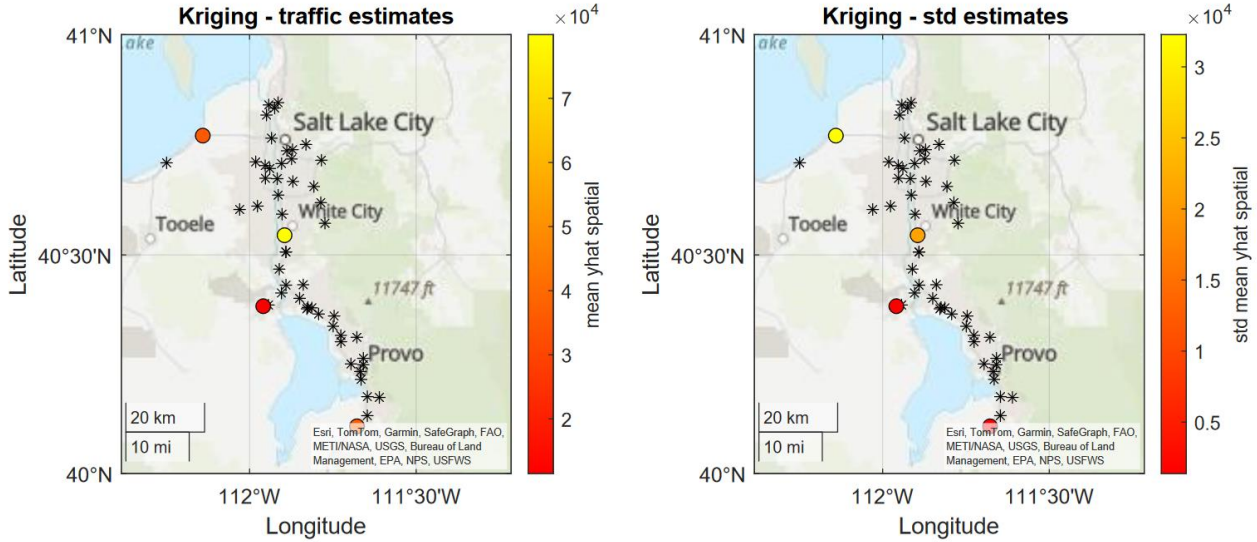


Figure 5: (Left) All the stations included in our dataset. (Right) Clustering technique applied on the 50 stations. The validation performances of these stations using this technique are listed in table 1

Figure 6: Kriging estimates are represented by the marked dots at specific spatial locations, while the black markers denote all the traffic stations included in the dataset



Seasonality management in DCM models

Differencing method

- From the time series literature one way to treat seasonality is by **differencing** or **seasonal differencing**
- Both methods can help to reduce or eliminate the seasonality trend in the data
- **Objective:** apply the differencing method to our response variable and see the results
- We will create the `DCM_m4_seas` which have the same implementation structure as `DCM_m4`

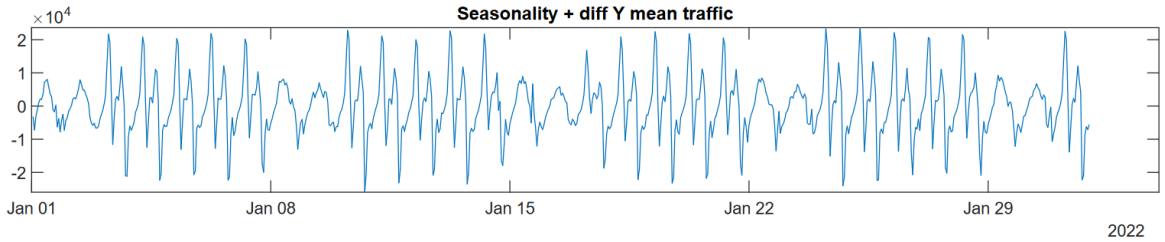


Figure 7: Temporal mean of the standardized traffic, computed after applying the differencing method to the response variable

Loading coefficient	Value	Std	$\|t\|$
weekend	-0.081	0.045	1.805
holidays	-0.095	0.084	1.138
mean temp	-0.127	0.025	5.083
mean prec	0.023	0.012	1.898
traffic on	0.286	0.026	11.018
hours	-0.095	0.011	8.351
interstate	-0.098	0.040	2.474
US	-0.097	0.033	2.969
RS	-0.097	0.032	3.019

Table 4: All β coefficients of the covariates included `DCM_m4_seas` model

Results

- The latent variable has the same form as the response variable
- Different temporal adjustment
- Latent variable is capturing a different phenomenon
- Residual distribution and correlation is better from before
- Differencing approach improved residual behaviour and overall statistics

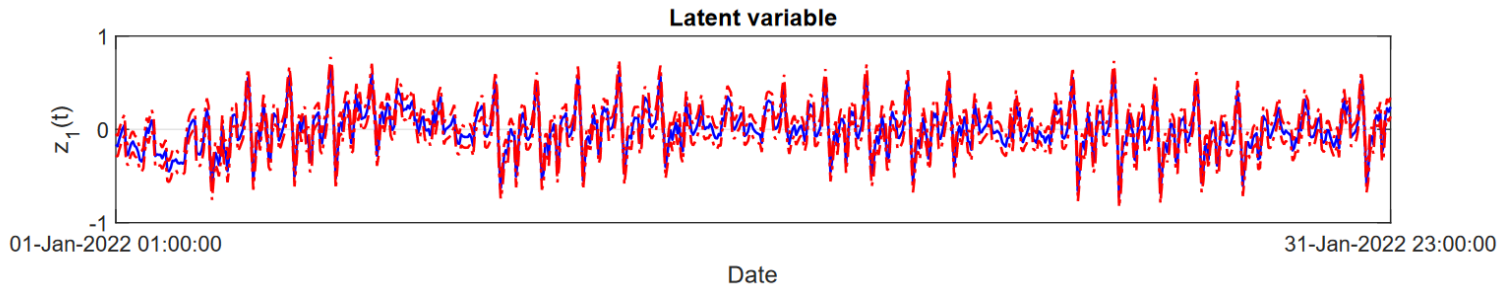


Figure 8: Latent variable estimated by the DCM_m4_seas model

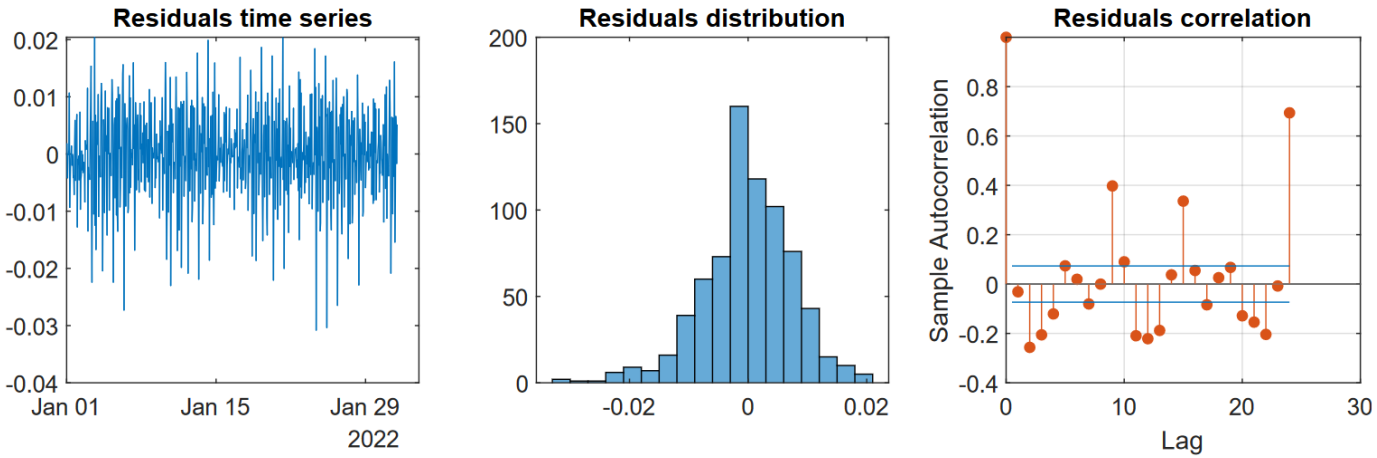


Figure 9: Mean temporal residuals statistics estimated from the `DCM_m4_seas` model in training phase

The HDGM

- The DCM tends to overfit at each time step
- HDGM overcomes the overfitting issue
- $\eta(\boldsymbol{s},t) \sim GP(0, \rho(||\boldsymbol{s} - \boldsymbol{s}'||; \boldsymbol{\theta}))$ is correlated over space but IID over time
- $z(\boldsymbol{s},t)$ is scalar and has Markovian dynamic
- α is a scale coefficient (v in D-STEM)
- g is the transition coefficient
- $\varepsilon(\boldsymbol{s},t) \sim N(0, \sigma_\varepsilon^2)$ is the measurement/model error
- The model parameter set is $\Psi = \{\boldsymbol{\beta}, \alpha, \sigma_\varepsilon^2, \boldsymbol{\theta}, g\}$

$$y(\boldsymbol{s},t) = \boldsymbol{x}_{\boldsymbol{\beta}}(\boldsymbol{s},t)'\boldsymbol{\beta} + \alpha z(\boldsymbol{s},t) + \varepsilon(\boldsymbol{s},t)$$

$$z(\boldsymbol{s},t) = g z(\boldsymbol{s},t-1) + \eta(\boldsymbol{s},t)$$

Formula: HDGM model description

HDGM modeling

HDGM implementation

- DCM you tend to overfit, because of no constraints on the spatially varying effects
- This lack of constraints is overcome with the HDGM with a more structured approach
- HDGM has same structure as the `DCM_m4`:
  - 100 EM iterations
  - Splitting into training and validation dataset
  - Passing all the covariates to the fixed effect component and a matrix of ones to the latent variable
- Only significant covariates are kept into the model: no overfitting, better generalization

Table 5: All β coefficients of the covariates included in the fixed effect part of the HDGM. Only the significative covariates are kept into the model to avoid overfitting. `mean temp`, as well as `mean prec` covariates are not significant so they were removed.

Loading coefficient	Value	Std	$\|t\|$
weekend	-0.268	0.034	7.943
holidays	-0.067	0.058	1.157
traffic on	0.780	0.013	57.987
hours	0.175	0.006	30.389
interstate	0.213	0.044	4.833
US	-0.893	0.065	13.809
RS	-0.836	0.046	18.371

Validation phase

- Highlighted spikes in the validation statistics indicate the model's incapacity to fully explain the total variance present in the data
- Poor estimation properties and inadequate generalization

- Sesonality pattern still present in the statistics
- Inability to accurately capture the high traffic phenomena during the week peak traffic times

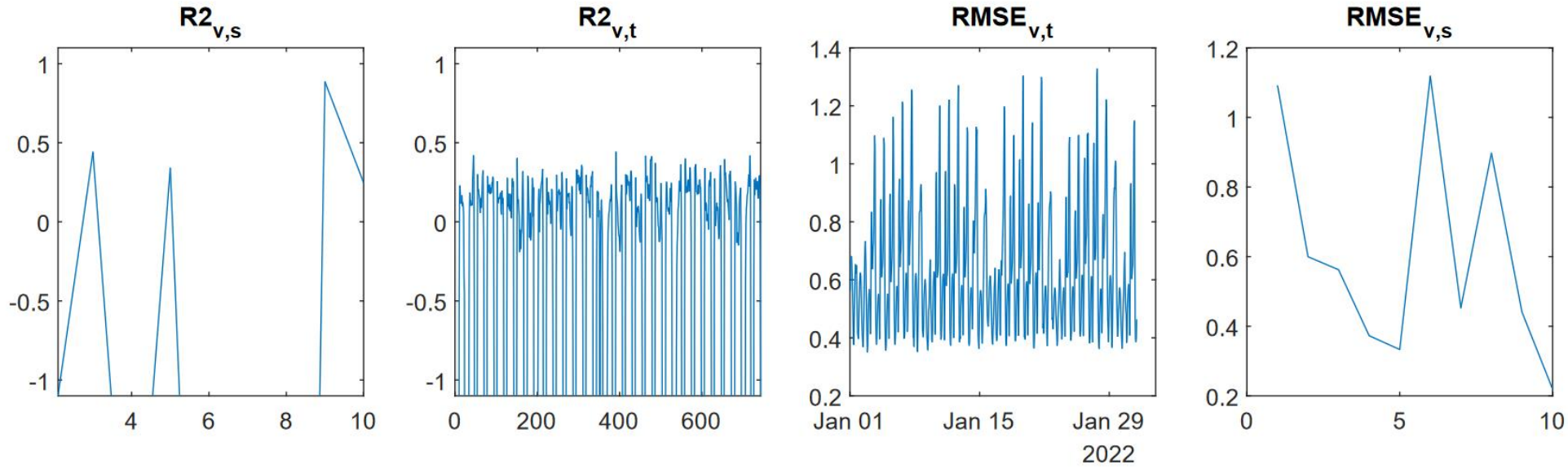


Figure 10: Temporal and spatial validation statics of the HDGM model. The spikes defined as peaks surpassing the imposed y-axis limit.

Validation dynamics on the territory

- Clustering method
- Training stations cover uniformly the entire area
- Validation stations trategically selected within the region

- Reduce uncertancy and increse estimates consistency
- Poor overall predictions by `HDGM` model

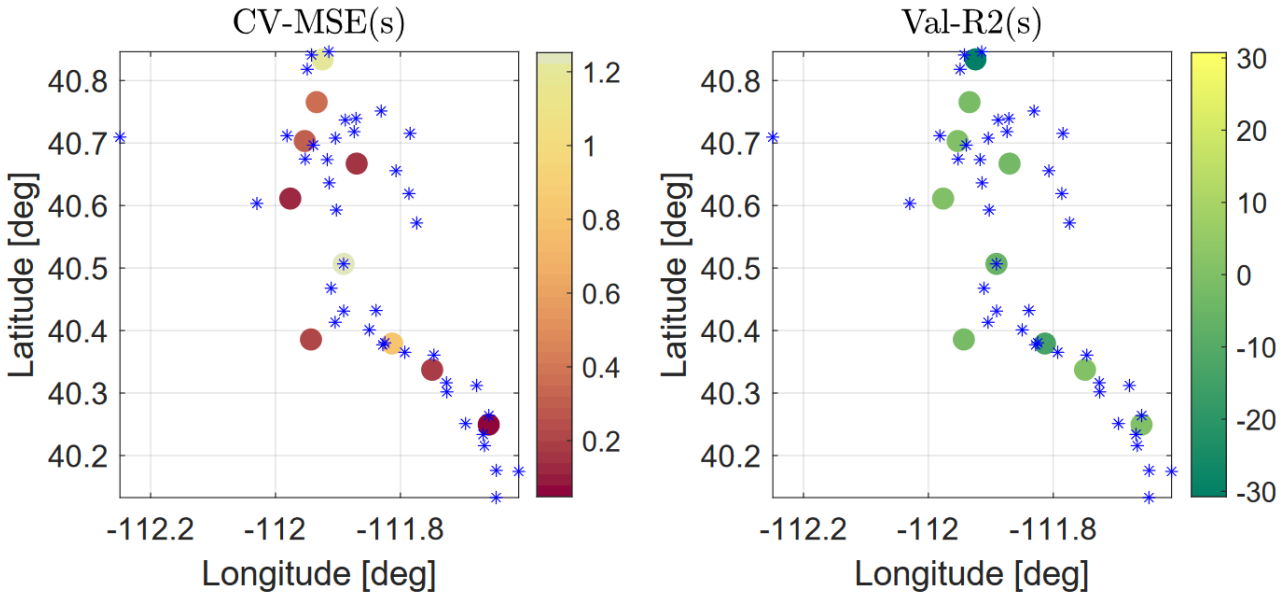


Figure 11: Validation statistics displayed spatially on the Salt Lake territory. (Blue markers) are the training stations instead the (Highlighted dots) are the validation stations.

Latent variable

- Each station has a latent variable with markovian dynamics
- Most latent variables exhibit similar temporal patterns, instead some portrays distinct behaviors
- Potential heterogeneity across space or conditions
- Model fails to capture all the variability within the traffic data

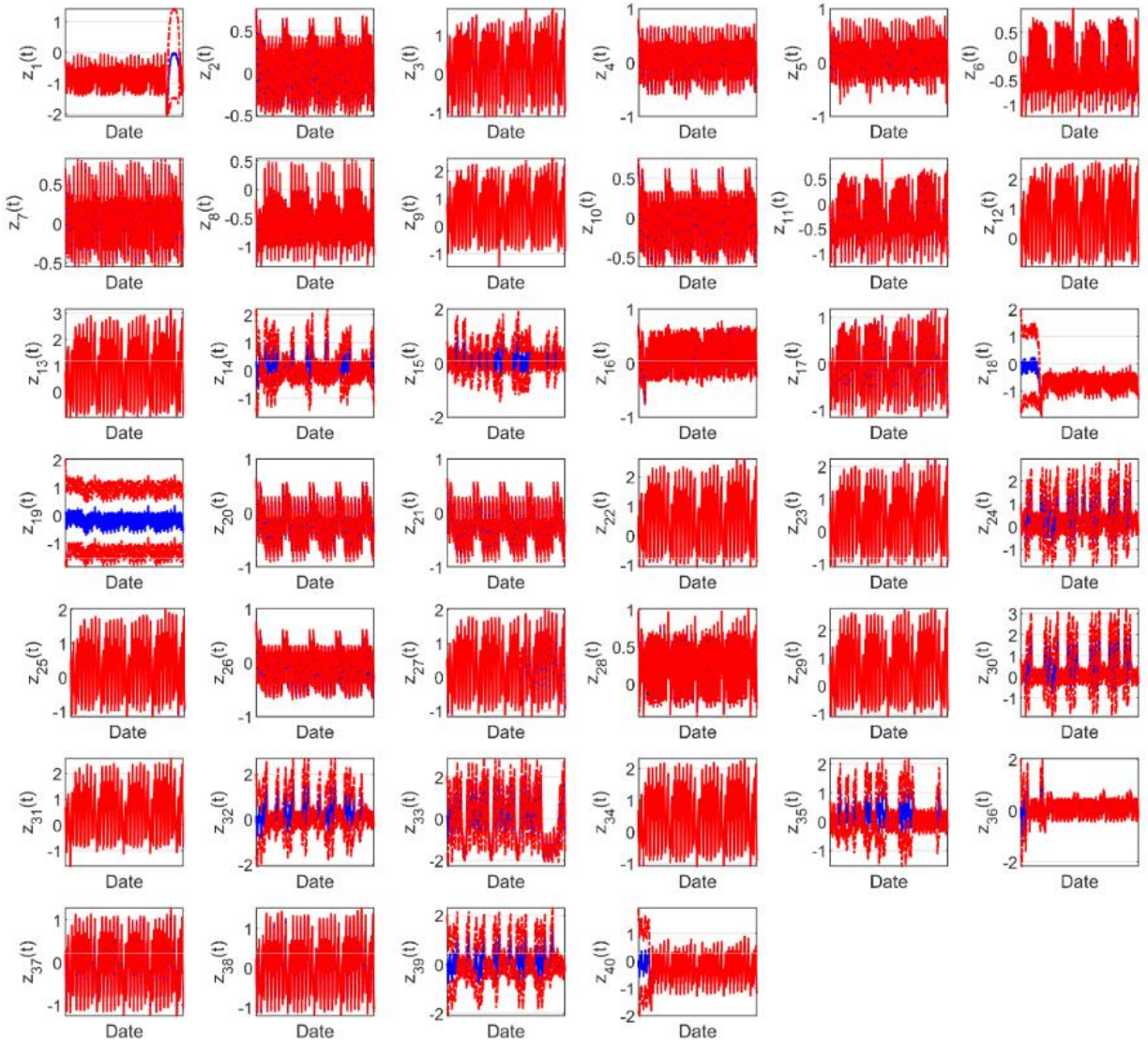


Figure 12: Latent variables estimated by the HDGM and their temporal behaviour per each station

Conclusions

- The traffic phenomenon is highly complex
- DCM and HDGM performances are low
- Theorical limitations or wrong enrichment ?
- Differencing method is worth the try
- `DCM_m4_seas` emerges as a more performant mode to apply to this specific case study

Models	$\overline{R}^2_{t,s}$	$\overline{R}^2_{v,s}$	$\overline{R}^2_{v,t}$	$\overline{RMSE}_{v,t}$	$\overline{RMSE}_{v,s}$	$logL$	$lbqtest$	$archtest$
DCM_m4	0.37	-2.76	0.44	0.39	0.34	1.10e+04	1	1
DCM_m4_seas	0.31	-0.28	0.35	0.32	0.35	1.12e+04	1	0
HDGM	0.97	-5.43	-14.11	0.63	0.60	2.19e+04	1	1
HDGM_seas	0.43	-2.61	-0.60	0.51	0.62	-1.57e+03	1	1

Table 6: Goodness statistics per each model analyzed, along with their residual tests

References

- **[1]** John Young Sorensen. Salt lake city traffic. https://www.kaggle.com/datasets/johnyoungsorensen/salt-lake-city-traffic, 2022.
- **[2]** Open-Meteo. Meteorological forecast service. https://www.open-meteo.com, 2024.
- **[3]** Francesco Finazzi and Alessandro Fassò. D-stem: A software for the analysis and mapping of environmental space-time variables. Journal of Statistical Software, 62, 12 2014. doi: 10.18637/jss.v062.i06.
- **[4]** Yaqiong Wang, Francesco Finazzi, and Alessandro Fass`o. D-stem v2: A software for modeling functional spatio-temporal data. Journal of Statistical Software, 99(10):1–29, 2021. doi: 10.18637/jss.v099.i10. URL https://www.jstatsoft.org/index.php/jss/article/view/v099i10.
- **[5]** Rob J Hyndman and George Athanasopoulos. Forecasting: principles and practice. OTexts, 2018.