Geo-Temporal Analysis on Salt Lake city traffic

Silviu Filote - 1059252
Jonathan Bommarito - 1068755

January 2024

Abstract

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.

Keywords: DCM, HDGM, multivariate space-time models, MATLAB, tuning procedures, seasonality, enrichment, validation, kaggle

1 Dataset description

The selected dataset is called "Salt Lake City Traffic" [1] and it's available on Kaggle site. The original dataset is divided in two different parts:

- "Utah Traffic": contains hourly traffic observations recorder by 50 stations, starting from 1 January 2022, 00:00 to 31 January 2022, 23:00. Missing data are identified as NaN values.
- "Utah Traffic MetaData": contains latitude and longitude information for each station and the route name where this one is placed.

Given the lack of additional usable variables as possible covariates, it was decided to address these missing pieces of information by enriching the dataset.

2 Dataset enrichment

The initial dataset provided has only one covariate, which is the route name where the station is located. To enrich the dataset, additional dummy variables were introduced. These include `weekend`, which indicates whether the current day is a Saturday or Sunday (set as 1 if it is, 0 otherwise), and `holidays`. Additionally, we introduced a variable to account the peak traffic hours called `traffic on`. The peak traffic hours can be seen as peaks in the response variable, so we set the value to 1 during the hours of 7 AM to 5 PM, where most of the traffic is recorderd and to 0 otherwise. Moreover, a variable for hours was introduced with the aim to explain the seasonal component through its inclusion in the analysis. Based on the route name information inside the initial dataset, we classified each station located at a specific route name as:

- `Interstate`: denoting high-speed roads spanning multiple states;
- `US`: representing U.S. Route national highways connecting cities and regions;
- `RS`: indicating State Route roads managed at the state level, often linking rural areas, cities, and other locales within a state.

These classifications were integrated into the dataset using dummy variables for each route type. In the end, the dataset was enriched with the precipitation and the temperature values for each single station at each time step using open-meteo.com API [2], which returns a time series by simply providing the coordinates, start date, end date, and frequency. Nevertheless, as the stations were close to each other, we observed their values were similar, indicating low spatial dependency. Consequently, we decided to evaluate the mean for all the stations and keep in the dataset just the mean precipitation (`mean prec`) and the mean temperature (`mean temp`) at each time step.

3 Research focus and methodology applied

The objective of this study is to examine the traffic as response variable of Salt Lake City through different geostatistical approaches, which involved spatio-temporal models application. The employed models are:

- Univariate - Dynamic Coregionalization Model (DCM)
- Univariate - Hidden Dynamic Geostatistical Model (HDGM)

A brief description of these models will be release in the following sections but both are treated and described thoroughly in the "D-STEM: A Software for the Analysis and Mapping of Environmental Space-Time Variables"[3]. These models will use initially all the covariates in order to explain the traffic amount and then the least significant ones will be removed. To test the performance of the models we evaluated:

- the residuals distribution, looking for eventual correlation over time. This analysis aims to uncover patterns and relationships left inside the residuals
- the estimates of the validation phase. The stations were divided into 10 spatial zones, and for each zone, a cluster was created, comprising stations that are nearby. Consequently, the validation set consisted of stations (totaling 10 stations, one per cluster) close to those used in the training set, thereby providing more consistent estimates. The validation processes yielded $RMSE_{v,t}$, $RMSE_{v,s}$, $R^2_{v,t}$, and $R^2_{v,t}$, which are crucial for evaluating the model's performance and generalization capability.
- log-likelihood model's parameter, which a higher value of this parameter grants a more accurate model parameters estimate as well as appropriate statistical inference.
- kriging performance, i.e. we will predict the traffic at unknown spatial locations using the model estimated along with the enrichment covariates for those points to see the output performances

Tuning approaches such as maximization of the log-likelihood function and EM convergence based on randomization of initial parameters were tested to enhance model performance and accuracy. All analyses conducted and model implementations in this paper are performed using the D-STEM v2 MATLAB library [4]. D-STEM uses the EM algorithm to estimate the model's parameters, but they required initialization values.

4 Univariate DCM

The first model applied to the traffic analysis is the univariate DCM (Dynamic Coregionalization Model) which is treated and implemented in the D-STEM software.

4.1 Model description

$$y(\boldsymbol{s},t) = \boldsymbol{x_\beta}(\boldsymbol{s},t)'\boldsymbol{\beta} + \boldsymbol{x_z}(\boldsymbol{s})'\boldsymbol{z}(t) + \omega(\boldsymbol{s},t) + \varepsilon(\boldsymbol{s},t) \tag{1}$$

$$\boldsymbol{z}(t) = \boldsymbol{G}\boldsymbol{z}(t-1) + \boldsymbol{\eta}(t) \tag{2}$$

$$\omega(\boldsymbol{s},t) = \sum_{j=1}^{c} \alpha_j x_j(\boldsymbol{s},t)\omega_j(\boldsymbol{s},t) \tag{3}$$

Remarks:

- $y(\boldsymbol{s},t)$ denotes the response variable representing the traffic captured by stations at spatial location $\boldsymbol{s}$ and time t, with the possibility of containing missing values.
- $\boldsymbol{x_\beta}(\boldsymbol{s},t)$ and $\boldsymbol{x_z}(\boldsymbol{s})$ are vectors of covariates with no missing values. The covariates in $\boldsymbol{x_\beta}(\boldsymbol{s},t)'\boldsymbol{\beta}$ component are the fixed effect in space and time of the response variable, whereas $\boldsymbol{x_z}(\boldsymbol{s})$ contains only spatial covariates which influence changes over time. This can be seen as a temporal correction made by the Markovian process over the fixed effect.
- Latent variable $w(\boldsymbol{s},t) \sim GP\big(0, \rho(||\boldsymbol{s}-\boldsymbol{s}'||;\boldsymbol{\theta})\big)$, correlated in space and IID over time. It can be seen as a spatial correction based on the correlation function over the fixed effect.
- $\boldsymbol{z}(t)$ is $q \times 1$ dimensional with Markovian dynamics

- $\boldsymbol{G}$ is a stable $q \times q$ transition matrix, it can be full, diagonal
- $\boldsymbol{\eta}(t) \sim N(\mathbf{0}, \Sigma_\eta)$ is the innovation with Σ_η the variance-covariance matrix
- $\varepsilon(\boldsymbol{s}, t) \sim N(0, \sigma^2_\varepsilon)$ is the measurement error
- The EM parameter vector is $\Psi = \{\boldsymbol{\beta}, \alpha, \sigma^2_\varepsilon, \boldsymbol{\theta}, \boldsymbol{G}, \boldsymbol{\Sigma}_\eta\}$, where initialization is required

4.2 Model parametrizations

The primary goal behind modeling the response variable with a DCM is to explore and inspect the variety of possible parameterizations of the models we can construct. Subsequently, the aim is to select the model which has the best performance overall.

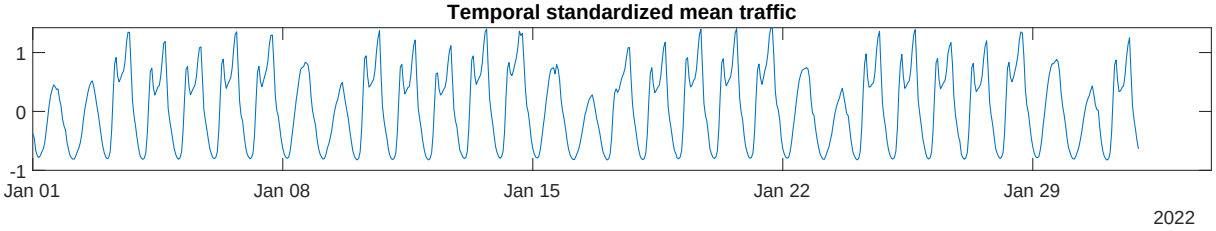

Figure 1: Temporal mean of the standardized traffic passed as response variable to the D-STEM software. We will refer to this variable as Y_t, which represents a time series created by averaging the hourly data from all available stations. Futhermore, $\overline{Y}_t$ is the average of Y_t.

Based on the seasonal response variable and the enrichment covariates, we systematically tested various configurations by adjusting the combination of covariates in each part of the DCM model. Table 1 presents only a subset of the models examined, aiming to highlight the most effective ones rather than listing all analyzed configurations. Each of the following models shares the same initial settings: 100 EM iterations and same starting parameter vector $\Psi = \{\boldsymbol{\beta}, \alpha, \sigma^2_\varepsilon, \boldsymbol{\theta}, \boldsymbol{G}, \boldsymbol{\Sigma}_\eta\}$.

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 log-likelihood parameter and residual tests. $\overline{R}^2_{t,s}$ is read as the spatial mean in training, so basically the first subscript indicates if the statistic is estimated in training or validation and the second one indicates if the statistic is spatial or temporal. *lbqtest* determines whether the mean residual distribution exhibits autocorrelation, while *archtest* evaluates the presence of conditional heteroscedasticity

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

According to training and validation statistics the `DCM_m4` seems to be a good overall model with good performance in training phase and the best ones in validation. Furthermore, the $logL$ is quite high aswell, granting in this way more consistent estimates, but the mean residuals distribution of the `DCM_m4` model exhibits autocorrelation and heteroscedasticity. The `DCM_m4` will be analized more in deapth in next section.

4.3 Model implementation

We splitted the total stations into training and valdation sets, where 40 stations were used in training and 10 stations in validation. To choose the validation stations we implemented a clustering technique, which is explained in the previous chapter. The implementation of `DCM_m4` in training phase includes all enrichment covariates with varying degrees of significance in explaining the response variable. Although `holidays`, `mean temp` and `mean prec` are considered the least impactful covariates, we will include them in the model based on their $|t|$ statistic. The effective construction of the `DCM_m4` model, according to the D-STEM v2 classes, can be summerized as:

- $\boldsymbol{x_\beta}(\boldsymbol{s}, t)$: loading coefficients related to $\boldsymbol{\beta}$, contains all the covariates
- $\boldsymbol{x_z}(\boldsymbol{s})$: loading coefficients related to $z(t)$, consists of a matrix of ones
- $\boldsymbol{x_p}(\boldsymbol{s})$: the loading coefficients $x_j(\boldsymbol{s}, t)$ for the latent spatial variables $\omega_j(\boldsymbol{s}, t)$, incorporates only `interstate`. As a result, we designate $x_p(\boldsymbol{s})$ as spatial dependency only.

Table 2: All covariates included in the $\boldsymbol{x_\beta}(\boldsymbol{s}, t)'\boldsymbol{\beta}$ part of the DCM model are listed in the following table. This component, $\boldsymbol{x_\beta}(\boldsymbol{s}, t)'\boldsymbol{\beta}$, characterizes the fixed effects introduced by covariates in explaining the response variable across temporal and spatial dimensions.

Loading coefficient	**Value**	**Std**	$\lvert t \rvert$
`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

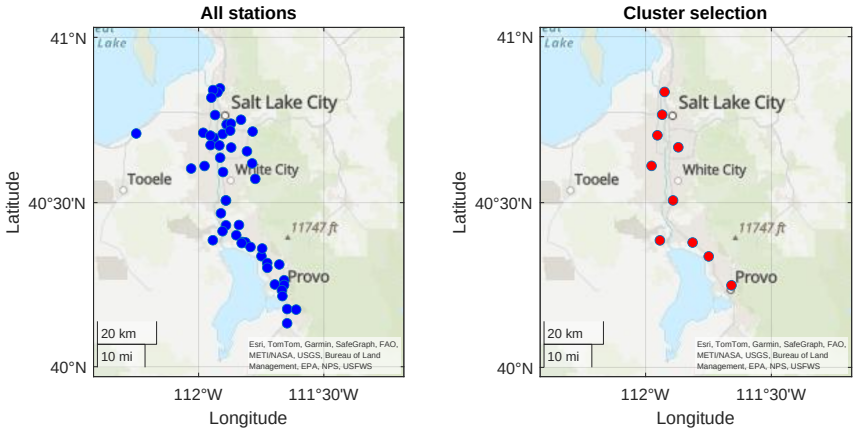

Figure 2: (Left) All the stations included in our dataset. (Right) Clustering technique applied on the 50 stations. The validation performances of these stations using this technique are listed in table 1

Reviewing the temporal validation statistics, we can see some strange behaviour that happens to the `DCM_m4` model. The current highlighted spikes in the R^2 parameters indicate the model's incapacity to fully explain the total variance present in the data. Consequently, this inadequacy portrays a poor estimation by the model. Moreover, the temporal behaviour of $RMSE_{v,t}$ reflects the same seasonality pattern of the response variable. This discrepancy arises from the `DCM_m4` model's inability to accurately capture the high traffic phenomena during the week peak traffic times.

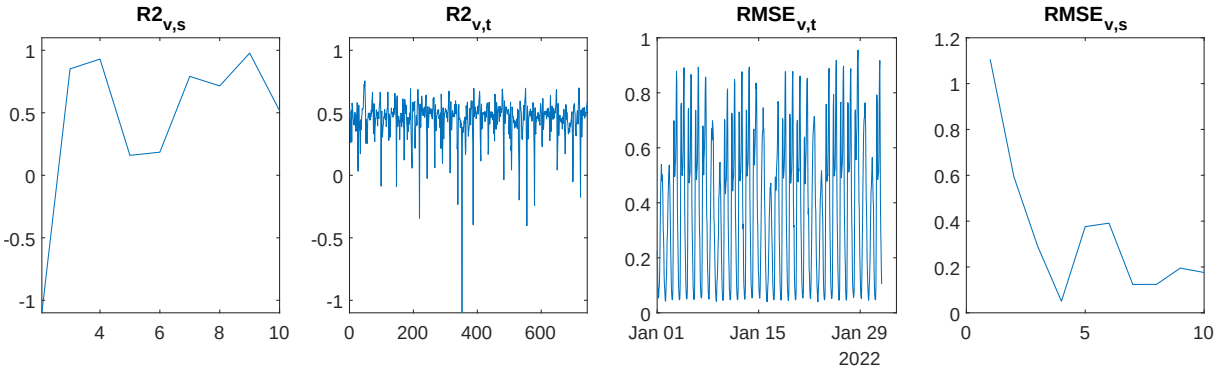

Figure 3: Temporal and spatial validation statistics of the `DCM_m4` model. The spikes defined as peaks surpassing the imposed y-axis limit.

The latent variable $z_1(t)$ can be seen as a temporal correction over the fixed effect of the model made by the covariates included into the $\boldsymbol{x_z}(\boldsymbol{s})'\boldsymbol{z}(t)$ part of the DCM model. Given the dynamic nature of the response variable, the latent variable exhibits an expected seasonal pattern. However, the seasonal adjustment made by the latent variable does not appear to be sufficient to eliminate the seasonal trend in the validation statistics and residual distributions.

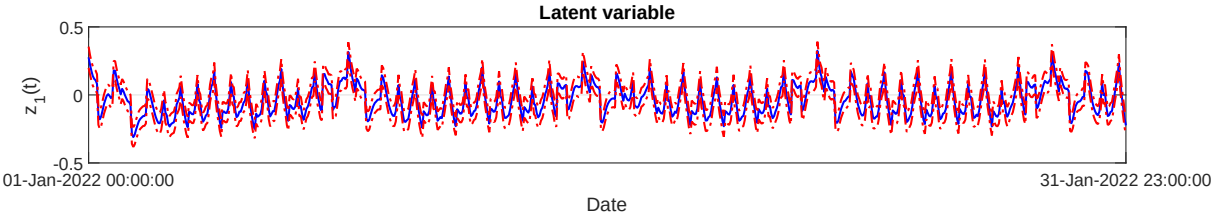

Figure 4: Latent variable estimated by the `DCM_m4` model

The test conducted on the mean temporal residual distribution from the training phase confirms heteroscedastic behavior and correlation over the 24-hour period, suggesting that the theoretical assumptions made in the DCM description model chapter regarding the residuals are not met. Additionally, there is an evident seasonal pattern remaining within the residuals that the model is not able to capture. Therefore, the `DCM_m4` fails to capture all the variability present in the initial data, as evidenced by the residual attitude.

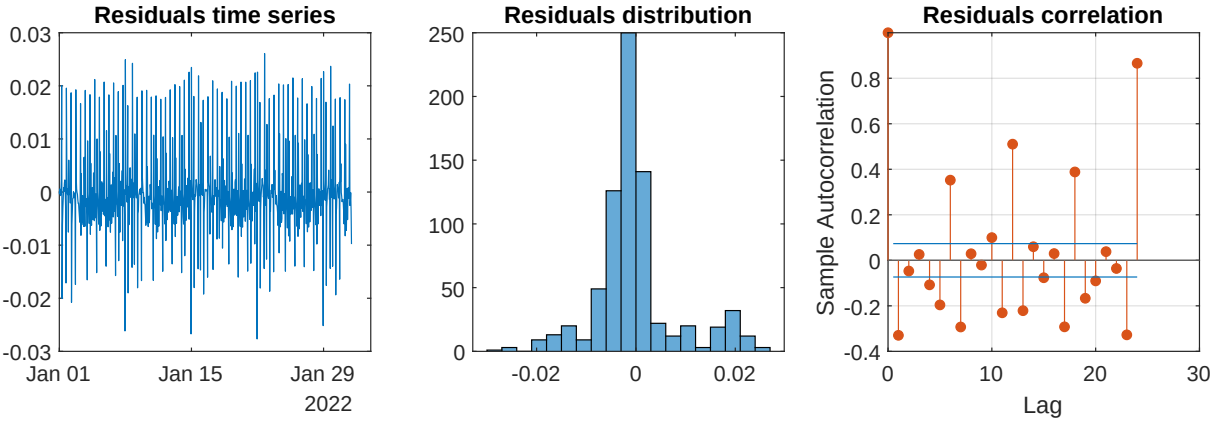

Figure 5: Temporal mean residual statistics derived from the training phase of the `DCM_m4` model. During the training phase, 80% of the available data is used, which includes 40 stations out of 50

To assess the predictive capability of the `DCM_m4` model, we utilized it to forecast traffic in various coordinates within the city where no monitoring stations were situated. These coordinates were chosen near our existing traffic network to maintain consistent estimates with minimal uncertainty. This operational approach is frequently referred to in geostatistical literature as kriging, which emphasizes the interpolation of data points to estimate values at unsampled locations. To increase the accuracy of the predictions at these spatial points, we also incorporated the enrichment covariates.

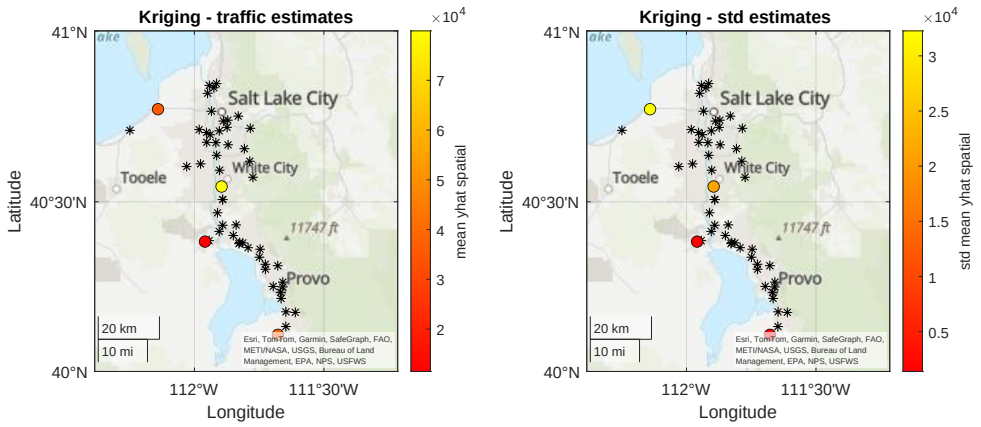

Figure 6: Kriging estimates are represented by the dots marker at specific spatial locations, while the black markers denote all the traffic stations included in the dataset

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

The mean uncertainty of $\overline{\sigma}_k$ is $1.4126e+04$, while the mean traffic estimated from 1 January 2022, 00:00, to 31 January 2022, 23:00 in the kriging locations is about $4.1284e+04$. These evaluations suggest a considerable level of uncertainty compared to the estimated traffic volume.

5 Seasonality management in DCM models

In time series literature one, way to treat the seasonality is by applying the **differencing** aproach. Basically, differencing consists in computing the differences between consecutive observations obtaining in this way a new time series $y_t^{'}$:

$$y_t^{'} = y_t - y_{t-1} \tag{4}$$

When differences are calculated between an observation and the previous observation from the same season, this is referred to as **seasonal differencing**:

$$y_t^{'} = y_t - y_{t-m}, \qquad m = the\ number\ of\ seasons \tag{5}$$

According to Rob J. Hyndman and George Athanasopoulos [5], these methods can help to reduce or eliminate the seasonality trend in the data obtaining a stationary time series. Both methods can be applied, and the execution order does not matter. In light of the above considerations, we have applied the differencing approach to our traffic dataset to reduce or even eliminate the seasonality pattern, and to assess the performance of the DCM model.

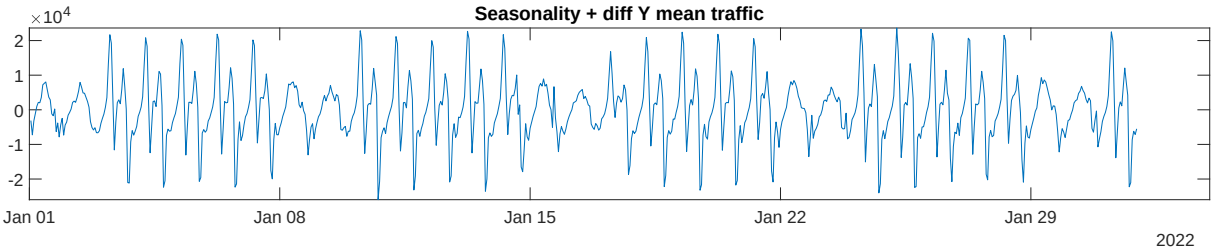

Figure 7: Temporal mean of the standardized traffic, computed after applying the differencing method to the response variable

The created model maintains the same implementation structure as `DCM_m4`, but with the differencing approach applied to the data. In fact, we will refer to this model as `DCM_m4_seas` to denote the method employed. According to the model result displayed in table 4, it is observable that the overall significance of the covariates has decreased drastically compared to the `DCM_m4`, except for the `mean temp` covariate which significance has increased. Furthermore, in `DCM_m4`, hours and traffic covariates have a positive impact on the response variable, while in `DCM_m4_seas`, we observe the opposite effect. The log-likehood, which is a statistic used to evaluate the good performance of the models, have more or less the same value in both models.

Table 4: All $\boldsymbol{\beta}$ coefficients of the covariates included in the $\boldsymbol{x}_{\boldsymbol{\beta}}(\boldsymbol{s}, t)$ part of the DCM, which purpose is to effectively manage the seasonality within the data

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

The latent variable measured has the same form as the temporal mean of the standardized traffic represented in figure 7, for a better understanding the Y_t after the differencing. This temporal adjustment is different from before because it capture a brand new phenomenon created by computing the differencing.

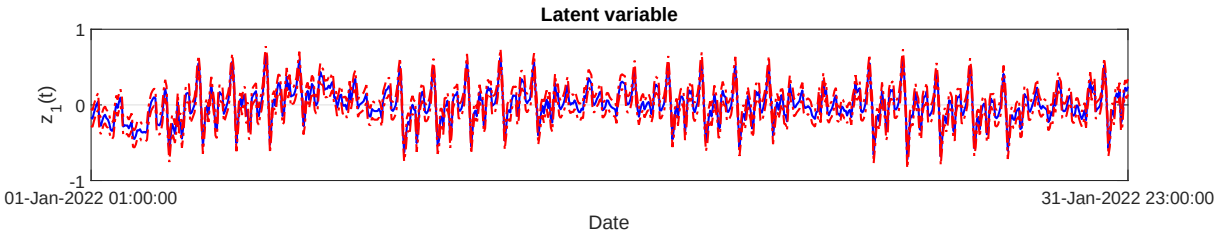

Figure 8: Latent variable estimated by the `DCM_m4_seas` model

The differencing approach has a significant impact on the residuals, which now show a more normal distribution and clearer evidence of reduced correlation among them compared to those in the `DCM_m4`.

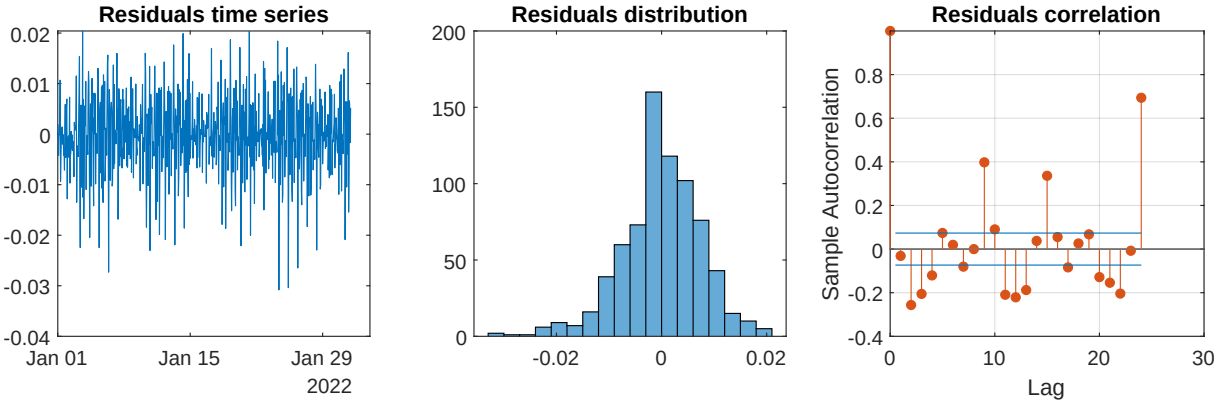

Figure 9: Mean temporal residuals statistics estimated from the `DCM_m4_seas` model in training phase

From the above considerations, it is evident that the differencing approach has stabilized the distribution of residuals and reduced autocorrelation without removing the seasonal component. Further in-depth analyses would be required, but the method itself remains highly valid for the specific application context in which we are operating.

6 Univariate HDGM

The DCM tends to overfit at each time step. This happens because the model is capable of fitting spatially independent patterns without considering previous time steps, due to the absence of constraints on spatially varying effects. The absence of constraints can lead to overly complex models that capture noise instead of true patterns. Consequently, the DCM often exhibits high training R^2 values but performs poorly in validation. This behaviour is evident in the table 1.
To overcome the overfitting issue, the HDGM (Hidden Dynamic Geostatistical Model) model offers a more structured and regularized approach by imposing a certain dynamics to the latent variable $z(\boldsymbol{s}, t)$.

6.1 Model description

$$y(\boldsymbol{s},t) = \boldsymbol{x_\beta}(\boldsymbol{s},t)'\boldsymbol{\beta} + \alpha z(\boldsymbol{s},t) + \varepsilon(\boldsymbol{s},t) \tag{6}$$

$$z(\boldsymbol{s},t) = gz(\boldsymbol{s},t-1) + \eta(\boldsymbol{s},t) \tag{7}$$

Remarks:

- $\eta(\boldsymbol{s},t) \sim GP\big(0, \rho(||\boldsymbol{s}-\boldsymbol{s}'||;\boldsymbol{\theta})\big)$ is correlated over space but IID over time
- The latent variable $z(\boldsymbol{s},t)$ now depends on $\boldsymbol{s}$. In a more detailed examination, exists a latent variable for each station in the dataset, which has a 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 EM model parameter set is $\Psi = \{\boldsymbol{\beta}, \alpha, \sigma_\varepsilon^2, \boldsymbol{\theta}, g\}$

6.2 Model implementation

The initialization parameters and implementation of the HDGM are identical to those of the previous models: 100 EM iterations, dataset splitting into training and validation sets, inclusion of all covariates in the $\boldsymbol{x_\beta}(\boldsymbol{s},t)$ fixed component, and assignment of a matrix of ones to the latent variable.

Table 5: All $\boldsymbol{\beta}$ coefficients of the covariates included in the $\boldsymbol{x_\beta}(\boldsymbol{s},t)$ part of the HDGM. Only the significative covariates are kept into the model in order to avoid overfitting

Loading coefficient	**Value**	**Std**	$\lvert t\rvert$
`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

In the initial analysis, both the `mean temp` and `mean prec` covariates were found to be insignificant. As a result, they will be removed to improve the generalization of the model estimates. Unfortunately, validation statistics fail to account for the seasonality present in the traffic data, leading to poor accuracy in estimating fluctuating patterns and, consequently, inadequate generalization.

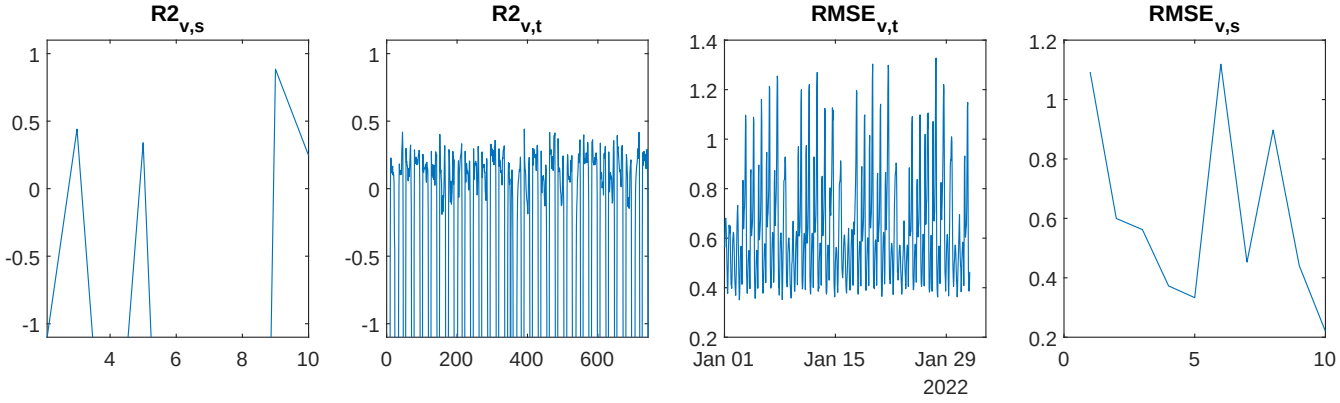

Figure 10: Temporal and spatial validation statics of the `HDGM` model. The spikes defined as peaks surpassing the imposed y-axis limit.

The training stations uniformly cover the entire spatial area. Employing the clustering technique, validation stations are strategically selected within regions already enclosed by the training stations. This strategy reduce the uncertainty all around the validation stations and increase the estimates' consistency. The estimates predicted in validation by the model are not promising, and the overall accuracy is poor, even when using the clustering approach.

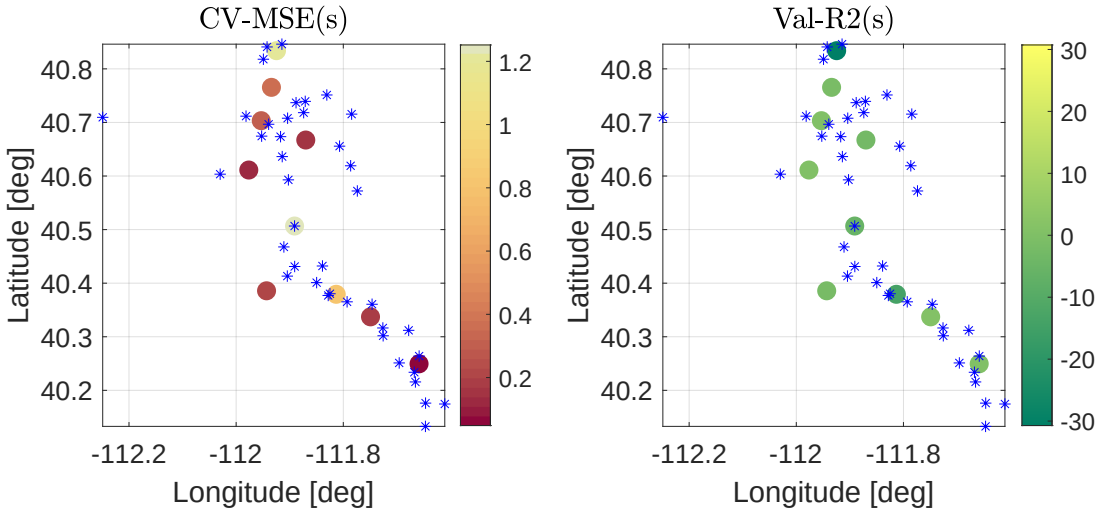

Figure 11: Validation statistics displayed spatially on the Salt Lake territory. (Blue markers) are the training stations instead the (Highlighted dots) are the validation stations.

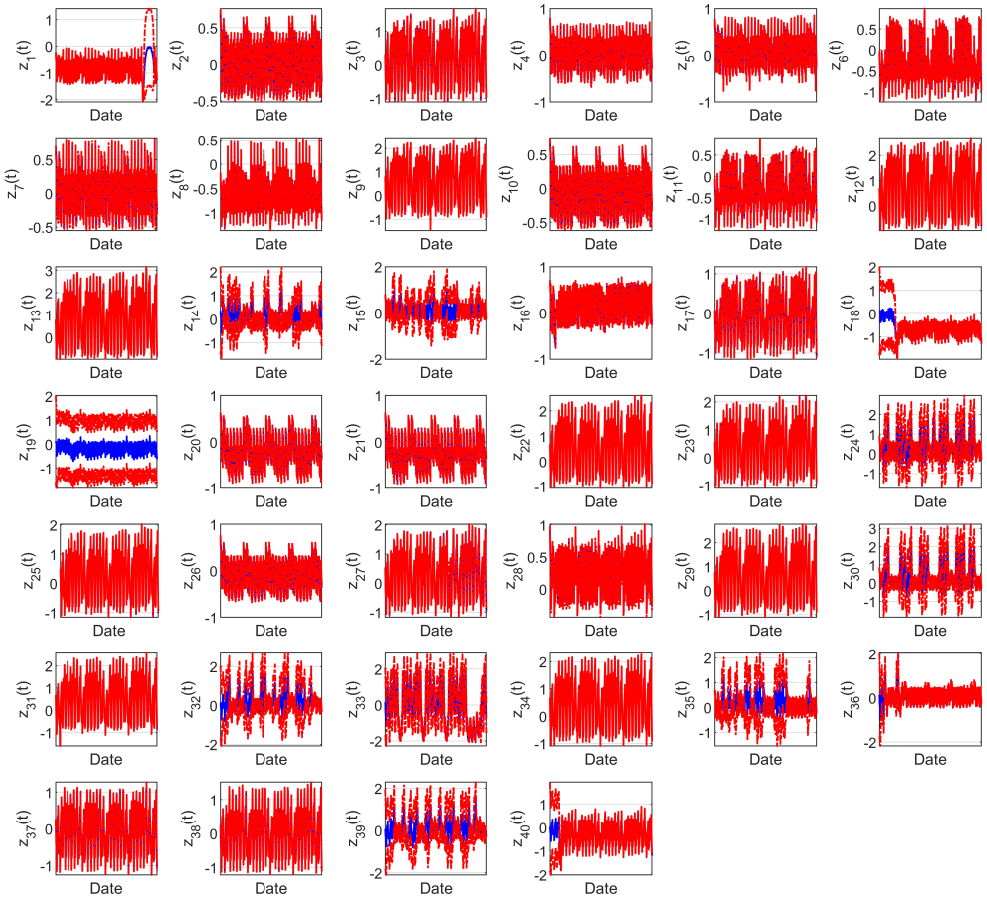

Figure 12: Latent variables estimated by the HDGM and their temporal behaviour per each station starting from 1 January 2022, 00:00 to 31 January 2022, 23:00.

Most of the latent variables displayed in figure 12 exhibit similar temporal patterns, with some stations portraying distinct behaviours. This observation suggests underlying patterns and variations within the dataset, indicating potential heterogeneity across different locations or conditions. The model fails to capture all the variability within the traffic data and this is accurately pictured in the residual statistics in figure 13, in fact: autocorrelation, seasonality, heteroscedasticity and deviation from normal distribution observed in the residuals are indicators of inadequate modelling.

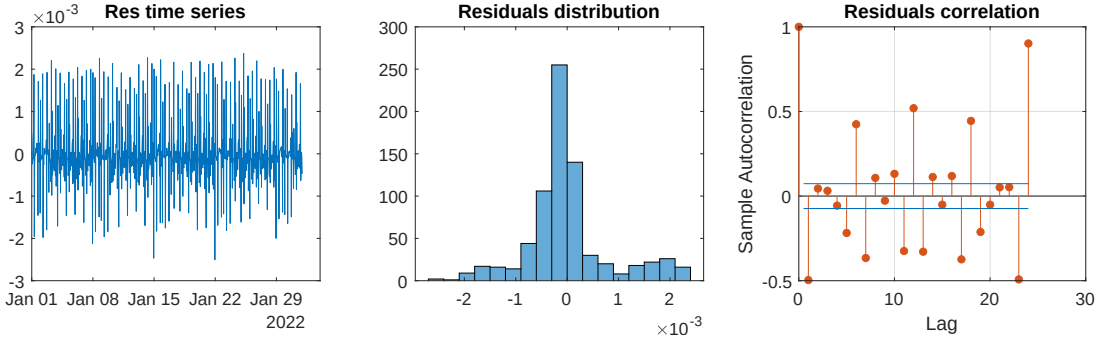

Figure 13: Mean temporal residuals statistics estimated from the `HDGM` model in training phase

As we did for the `DCM_m4` we tried to implement the `HDGM_seas` which purpose is to mitigate the seasonality of the traffic data with the intention of getting better residuals behaviour and model estimates, but in this case that doesn't happen so we are not going to add the specific deploying section of the `HDGM_seas` model.

7 Conclusions

The traffic phenomenon is highly complex, and modelling it using `DCM` and `HDGM` may not fully explain this variability. Similar outcomes are observed when employing the differencing method, which also fails in mitigating the seasonality effect of the traffic and also getting the worst estimation capabilities in some cases. We remain uncertain whether this limitation results from the theoretical modelling of the models deployed or the absence of more significant covariates that were not captured during the enrichment phase. However, upon comparing the analysed models based on the most critical statistics, it appears that `DCM_m4_seas` emerges as a more performant model for this particular case study, in despite of the modest statistics observed during the training process.

Table 6: Goodness statistics for all analyzed models, along with their residual tests

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

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 Fassò. 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.

List of Figures

List of Tables