Less But Better Towards better AQ Monitoring by learning Inducing Points for Multi-Task Gaussian Processes

Progyan Das PROGYAN.DAS@IITGN.AC.IN
Department of Computer Science and Engineeing
IIT Gandhinagar
Gujarat, India

Mihir Agarwal AGARWALMIHIR@IITGN.AC.IN
Department of Electrical Engineering
IIT Gandhinagar
Gujarat, India

Abstract

Air pollution is a pressing global issue affecting both human health and environmental sustainability. The high financial burden of conventional Air Quality (AQ) monitoring stations and their sparse spatial distribution necessitate advanced inferencing techniques for effective regulation and public health policies. We introduce a comprehensive framework employing Variational Multi-Output Gaussian Processes (VMOGP) with a Spectral Mixture (SM) kernel designed to model and predict multiple AQ indicators, particularly $PM_{2.5}$ and Carbon Monoxide (CO). Our method unifies the strengths of Multi-Output Gaussian Processes (MOGPs) and Variational Multi-Task Gaussian Processes (VMTGP) to capture intricate spatio-temporal correlations among air pollutants, thus delivering enhanced robustness and accuracy over Single-Output Gaussian Processes (SOGPs) and state-of-the-art neural attention-based methods. Importantly, by analyzing the variational distribution of auxiliary inducing points, we identify high-information geographical locales for optimized AQ monitoring frameworks. Through extensive empirical evaluations, we demonstrate superior performance in both accuracy and uncertainty quantification. Our methodology promises significant implications for urban planning, adaptive station placement, and public health policy formulation.

Keywords: Gaussian Process, Explainability, Adaptive Deployment, Air Quality

1. Introduction

Air pollution poses a significant global health risk, making accurate and robust Air Quality (AQ) monitoring essential for public health policy formulation, environmental conservation, and effective mitigation strategies [12, 14, 15, 18]. The sparse distribution of existing AQ monitoring stations, coupled with the limitations of traditional interpolation techniques and physics-based models, constrains our ability to precisely capture the complex and interdependent dynamics of air pollutants [2, 7, 12]. Particularly concerning are fine particulate matter (PM2.5) and Carbon Monoxide (CO), which are primary indicators of air quality and have severe health implications [1, 21]. Their spatiotemporal dynamics in urban settings present a challenging prediction problem.

Multi-Output Gaussian Processes: Single-Output Gaussian Processes (SOGPs) have gained traction for their probabilistic predictions and uncertainty management [17, 20]. However, these models often ignore the potential temporal and spatial correlations among different pollutants. In contrast, Multi-Output Gaussian Processes (MOGPs) offer an advanced framework to model these correlations, thus enhancing the robustness and accuracy of AQ predictions [8]. MOGPs allow for a shared representation of the input space, enabling the model to exploit the inherent correlations between tasks like predicting PM2.5 and CO concentrations concurrently.

Variational Distribution: Another innovative aspect of our approach involves the analytical investigation of the variational distribution of auxiliary inducing points [19]. These inducing points are critical for the scalability and accuracy of Gaussian Process (GP) models. Our empirical results indicate a marked clustering of inducing points around specific monitoring stations, suggesting these locations contain higher informational content. Further validation confirms that these information-rich stations yield lower Root Mean Square Error (RMSE) when used individually for predictive tasks, thereby guiding optimized sensor deployment.

This paper aims to provide a comprehensive approach to AQ monitoring by synergistically combining the predictive power of MOGPs with the analytical rigor of variational inference, addressing both the challenges of accurate pollutant prediction and optimized sensor deployment.

2. Exploiting Spatio-Temporal Correlations:

The examination of the correlation between $PM_{2.5}$ and CO [22], [16] levels clearly foreshadow the enhancement in predictive accuracy when incorporated as inputs in the model, thereby highlighting a temporal linkage. Further, the scrutiny of spatial and temporal correlation through varying the number of monitoring stations revealed a decline in error rates with the utilization of multi-output Gaussian Processes (MOGPs) [13] alongside data from three stations as opposed to employing a single station with Gaussian Processes (GPs) or Random Forests (RFs)

Temporal Correlations and Mechanisms in Multi-Task Learning The correlations between $PM_{2.5}$ and CO serve as a critical focal point in our model, confirming that incorporating these temporal relationships enhances predictive accuracy [16, 22]. Particularly, data from three monitoring stations demonstrated marked improvements in error rates when employing MOGPs over traditional models like SOGPs or Random Forests (RFs) [13]. These observed correlations between CO and $PM_{2.5}$ are complex and possibly arise from multiple sources such as industrial activities, vehicle emissions, and natural phenomena [11, 23]. The data supports the notion that CO may even act as a precursor to $PM_{2.5}$ in some atmospheric chemical reactions [9, 24]. This multi-faceted relationship serves as a strong justification for adopting a multi-task learning approach, offering a richer perspective for AQ monitoring.

Our paper's primary contributions can be summarized in two overarching aspects:

- We propose a highly scalable and accurate MOGPs-Variational Inference framework that specializes in spatio-temporal AQ inference, particularly focusing on $PM_{2.5}$ and

CO. The model's efficiency and robustness are enhanced through an analytical treatment of Variational Inference for inducing points [19].

- By inspecting the variational distribution of the auxiliary set of inducing points, we discover information-rich geographical locales, thereby contributing to the development of adaptive AQ monitoring, by targeted deployment of additional AQ monitoring stations in under-sampled areas.

3. Methodology

Dataset The dataset harnessed in this study encompasses hourly measurements of $PM_{2.5}$ and CO from 36 monitoring stations scattered across Beijing, supplemented by meteorological data from the respective district, spanning the period from May 1, 2014, to April 30, 2015. We discuss the dataset in more detail in the appendix.

Our Approach The problem we address is concisely defined as the following,

> In our study, we aim to address two key objectives using a set of air quality (AQ) monitors $\mathcal{S}$, timestamps T, and pertinent features. First, we seek to forecast $PM_{2.5}$ and CO levels at new geographic locations $\mathcal{S}^*$ for the same time intervals T employing Multi-Output Gaussian Processes (MOGPs). Second, we utilize Variational Inference to optimize the deployment of new AQ monitoring stations by examining the spatial clustering of auxiliary inducing points in the model.

Traditional GPs model a single output, treating each output (here, pollutant) independently, thus overlooking potential correlations between different outputs, which are crucial for more accurate AQ inference. Our detailed experimentation confirms that our covariates and outputs exhibit significant spatial and temporal correlation.

Gaussian Process Models We investigate Gaussian Process (GP) methods for multi-output air pollution prediction, with a particular focus on Multi-Task Gaussian Processes based on the Linear Model of Coregionalization (LMC) and a non-stationary Spectral Mixture Kernel to capture complex temporal phenomena[4]. To mitigate computational burdens, we employ sparse versions of these models, using variational strategies for the effective selection of inducing points [10, 19], which we later exploit for adaptive AQ monitoring design.

Linear Model of Coregionalization (LMC) LMC allows joint analysis of pollutants via shared latent functions. It can be formally described as $g_k(x) = \sum_{d=1}^{D} B_{kd} f_d(x)$, where B is a coregionalization matrix.

Spectral Mixture (SM) Kernel and Multitask Kernel The SM kernel captures complex periodicity in AQ data with a Gaussian mixture model, formally given by $k_{SM}(x, x')$. We extend this to a multi-task scenario using a Kronecker product with a task covariance matrix K_{task}, yielding $k(x, x', k, k')$.

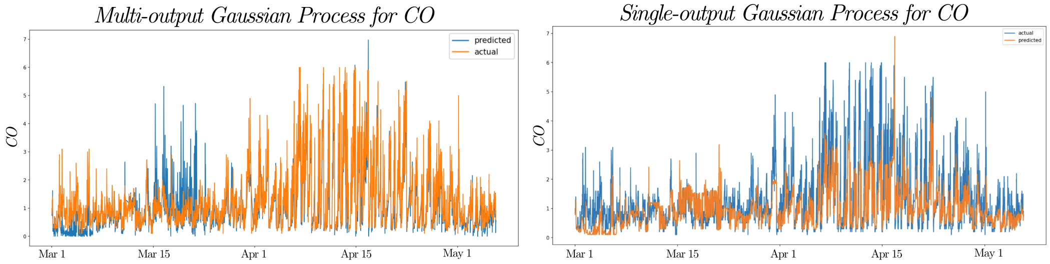

Figure 1: A comparison of CO prediction for MOGP vs. SOGP; as we can see, the MOGP does much better at predicting peaks in the data.

4. Experimentation and Evaluation

We conducted a series of experiments to evaluate the performance of our proposed MOGPs framework in comparison with other established models. The experimentation was done in the context of predicting $PM_{2.5}$ and CO concentrations, assessing both spatial and temporal correlations. The models compared include Multi-Output SGPR (MOSGPR), Random Forest (RF), K-Nearest Neighbors (KNN), Linear Regression (LR), Extreme Gradient Boosting (XGB), and Multi-Layer Perceptron (MLP). The empirical evaluations were conducted under diverse settings to explore temporal and spatial correlations between $PM_{2.5}$ and CO. The Root Mean Square Error (RMSE) was utilized as the performance metric in each case.

Independent Forecasting of $PM_{2.5}$ and CO A secondary line of investigation focused on the models' performance when incorporating the other pollutant as an input feature. The models were re-evaluated and the Mean Test Errors are summarized in Table **??** for CO and $PM_{2.5}$.

(a) CO Forecasting with $PM_{2.5}$ Input

Model	w/o $PM_{2.5}$	w/ $PM_{2.5}$
Exact SOGP	**0.55**	**0.41**
RF	0.53	0.43
KNN	0.54	0.56
XGB	0.56	0.43

(b) $PM_{2.5}$ Forecasting with CO Input

Model	w/o CO	w CO
Exact SOGP	**24.78**	**20.32**
RF	27.19	22.63
KNN	28.22	36.61
XGB	27.69	21.78

Table 1: RMSE on Independent forecasting of $PM_{2.5}$ and CO with each other in the input.

Multi-Task Forecasting of $PM_{2.5}$ and CO Initially, the models were employed to predict $PM_{2.5}$ and CO separately. The enlisted models include Multi-Output and Single-Output Exact Gaussian Process Regression (MOEGPR), Multi-Output and Single-Output Sparse Gaussian Process Regression (MOSGPR), as well as a baseline with [3], [5], Multi-Layer Perceptrons (MLP) both in multitask and single task settings. The obtained RMSE for $PM_{2.5}$ and CO are illustrated in Table 2.

The use of multiple monitoring stations, even without explicit location data, leads to lower RMSEs, confirming the value of spatial correlation. Lastly, our MOGP model with

Model	$PM_{2.5}$ (RMSE)	CO (RMSE)
Exact MOGP	**25.11**	**0.42**
Exact SOGP	*26.67*	*0.63*
Sparse MOGP (500 IP[1])	45.76	0.71
Sparse SOGP (500 IP[1])	47.27	0.81
KNN	38.64	0.65
LR	57.38	0.72
XGB	27.90	0.54
MLP	79.56	0.74

Table 2: RMSE on for multi-task forecasting based on five inputs – (latitude, longitude, temperature, humidity, wind speed). Best result is in **bold**, second-best is in *italics*.

LMC and SM Kernel outperforms both state-of-the-art SOGPs and neural-attention baselines, particularly when variational strategies for inducing points are employed.

Analysis of Variational Distribution and Locations By analyzing the inducing points sampled from the variational distribution, we can better understand the model's focus on certain air quality monitoring stations. We use the Haversine distance formula to account for the Earth's curvature when comparing locations. We find that some monitoring stations attract more inducing points, indicating these areas provide more useful information for the model. This helps in both improving the model's accuracy and in deciding where to best place additional sensors.

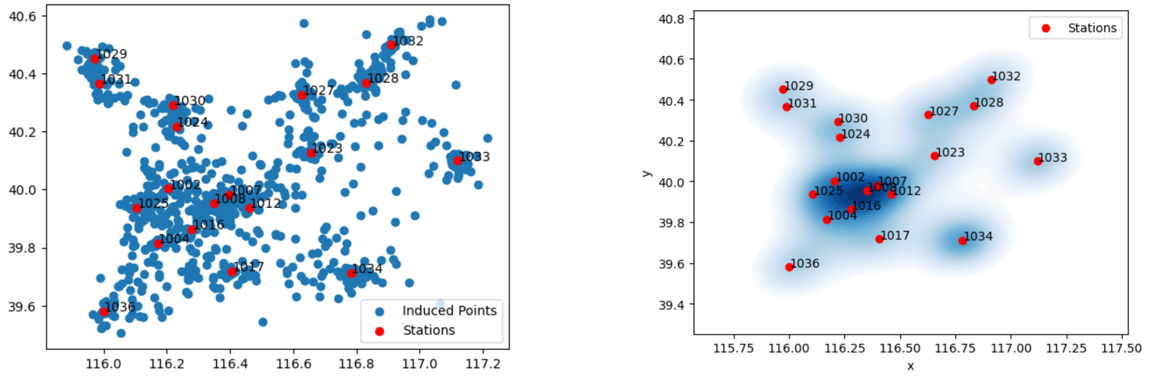

Figure 2: *Left*: The latitude/longitude indices of the inducing points sampled from the variational distribution plotted with the location of the AQ monitors. *Right*: A Kernel Distribution Estimation (KDE) plot of the inducing points, as an approximation of the variational distribution

5. Discussion

Our study on the variational distribution of auxiliary inducing points in the context of Variational Spectral Multi-Task Gaussian Processes has identified key geographical areas

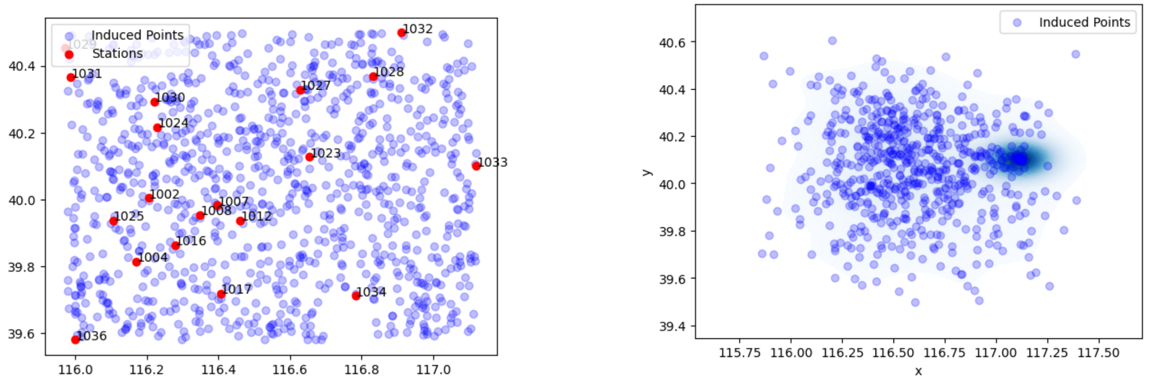

Figure 3: *Left*: The inducing point sampling for an untrained Gaussian Process. *Right*: KDE plot when the Gaussian process has been only trained on one station – the distribution peaks at that station.

that are high in information content. These areas are predominantly around specific air quality monitoring stations and serve as reliable sources for our model to acquire data and make accurate air quality predictions. On the flip side, this analysis also highlights areas that are currently under-sampled, providing lower informational value to the model.

To address this information imbalance, targeted deployment of additional AQ monitoring stations in these under-sampled areas could be a strategic move. This would not only improve the information density of our monitoring network but also enhance the model's ability to capture the complexities and variances in air pollution levels across different regions.

6. Conclusion

In this paper, we presented a comprehensive framework for air quality (AQ) monitoring that synergistically combines Multi-Output Gaussian Processes (MOGPs) with variational inference. Our methodology significantly advances the state-of-the-art in AQ monitoring by leveraging the temporal and spatial correlations among pollutants like PM2.5 and CO. The MOGPs model allows for shared representations of the input space, making it superior to existing Single-Output Gaussian Process (SOGP) models and traditional tree-based methods in both prediction accuracy and uncertainty quantification.

We introduced an analytical treatment of variational inference to identify the optimal positioning of auxiliary inducing points, thereby enhancing the model's scalability and performance. Our empirical results demonstrated that certain air quality monitoring stations possess higher informational value, revealing key regions where the variational distribution peaks. This, in turn, has substantial implications for the strategic deployment of AQ monitoring stations, particularly in regions that are currently under-sampled and offer lower information richness.

References

[1] Nicholas Apergis et al. "US state-level carbon dioxide emissions: does it affect health care expenditure?" In: *Renewable and Sustainable Energy Reviews* 91 (2018), pp. 521–530.

[2] Alexander Baklanov et al. "Potential and shortcomings of numerical weather prediction models in providing meteorological data for urban air pollution forecasting". In: *Water, Air and Soil Pollution: Focus* 2 (2002), pp. 43–60.

[3] Daniel Beck, Kashif Shah, and Lucia Specia. "Shef-lite 2.0: Sparse multi-task gaussian processes for translation quality estimation". In: *Proceedings of the Ninth Workshop on Statistical Machine Translation.* 2014, pp. 307–312.

[4] Gilles Bourgault and Denis Marcotte. "Multivariable variogram and its application to the linear model of coregionalization". In: *Mathematical Geology* 23 (1991), pp. 899–928.

[5] Tianqi Chen et al. "Xgboost: extreme gradient boosting". In: *R package version 0.4-2* 1.4 (2015), pp. 1–4.

[6] Weiyu Cheng et al. "A neural attention model for urban air quality inference: Learning the weights of monitoring stations". In: *Proceedings of the AAAI conference on artificial intelligence.* Vol. 32. 1. 2018.

[7] Seyyed Shahabaddin Hosseini Dehshiri and Bahar Firoozabadi. "A multi-objective framework to select numerical options in air quality prediction models: A case study on dust storm modeling". In: *Science of The Total Environment* 863 (2023), p. 160681.

[8] Oliver Hamelijnck et al. "Multi-resolution multi-task Gaussian processes". In: *Advances in Neural Information Processing Systems* 32 (2019).

[9] Jiming Hao and Litao Wang. "Improving urban air quality in China: Beijing case study". In: *Journal of the Air & Waste Management Association* 55.9 (2005), pp. 1298–1305.

[10] James Hensman, Alexander Matthews, and Zoubin Ghahramani. "Scalable variational Gaussian process classification". In: *Artificial Intelligence and Statistics.* PMLR. 2015, pp. 351–360.

[11] Peter Lenschow et al. "Some ideas about the sources of PM10". In: *Atmospheric environment* 35 (2001), S23–S33.

[12] Wenjun Li et al. "Air quality improvement in response to intensified control strategies in Beijing during 2013–2019". In: *Science of the Total Environment* 744 (2020), p. 140776.

[13] Haitao Liu, Jianfei Cai, and Yew-Soon Ong. "Remarks on multi-output Gaussian process regression". In: *Knowledge-Based Systems* 144 (2018), pp. 102–121.

[14] Gerardo Sanchez Martinez et al. "Health impacts and economic costs of air pollution in the metropolitan area of Skopje". In: *International journal of environmental research and public health* 15.4 (2018), p. 626.

[15] Manuel Méndez, Mercedes G Merayo, and Manuel Núñez. “Machine learning algorithms to forecast air quality: a survey”. In: *Artificial Intelligence Review* (2023), pp. 1–36.

[16] Gabriel Parra and Felipe Tobar. “Spectral mixture kernels for multi-output Gaussian processes”. In: *Advances in Neural Information Processing Systems* 30 (2017).

[17] Zeel B Patel et al. “Accurate and scalable gaussian processes for fine-grained air quality inference”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 11. 2022, pp. 12080–12088.

[18] Zongbo Shi et al. “Introduction to the special issue “In-depth study of air pollution sources and processes within Beijing and its surrounding region (APHH-Beijing)””. In: *Atmospheric Chemistry and Physics* 19.11 (2019), pp. 7519–7546.

[19] Michalis Titsias. “Variational learning of inducing variables in sparse Gaussian processes”. In: *Artificial intelligence and statistics*. PMLR. 2009, pp. 567–574.

[20] Peng Wang et al. “A Gaussian process method with uncertainty quantification for air quality monitoring”. In: *Atmosphere* 12.10 (2021), p. 1344.

[21] Jiansheng Wu et al. “Estimation of the PM 2.5 health effects in China during 2000–2011”. In: *Environmental Science and Pollution Research* 24 (2017), pp. 10695–10707.

[22] Yangyang Xie et al. “Spatiotemporal variations of PM2. 5 and PM10 concentrations between 31 Chinese cities and their relationships with SO2, NO2, CO and O3”. In: *Particuology* 20 (2015), pp. 141–149.

[23] Yang Yu et al. “Dynamics and origin of PM 2.5 during a three-year sampling period in Beijing, China”. In: *Journal of Environmental Monitoring* 13.2 (2011), pp. 334–346.

[24] Ying Zhou et al. “Temporal and spatial characteristics of ambient air quality in Beijing, China”. In: *Aerosol and Air Quality Research* 15.5 (2015), pp. 1868–1880.

Appendix A. Variational Multi-Task Gaussian Processes

In this appendix, we delve deeper into the rationale behind our choice of employing Variational Multi-Task Gaussian Processes (VMTGPs) for the inference of air-quality indicators. The unique characteristics of urban air quality data, encompassing spatial and temporal correlations between pollutants, demand a sophisticated modeling approach capable of capturing these intricacies. Our choice of VMTGP hinges on several compelling facets which render it particularly suited for air-quality inference.

A.1 Modeling Spatial and Temporal Correlations

The core essence of a Multi-Task Gaussian Process (MTGP) model lies in its ability to model correlations across different tasks, in our case, the prediction of $PM_{2.5}$ and CO concentrations. The temporal and spatial dependencies between these pollutants are crucial for accurate predictive modeling.

Multi-Task Gaussian Processes (MTGPs) Multi-Task Gaussian Processes extend traditional Gaussian Processes to multiple correlated tasks. Given T tasks and N observations, the covariance matrix of a MTGP is expressed as a Kronecker product of the task and input covariance matrices:

$$\mathbf{K} = \mathbf{K}_{task} \otimes \mathbf{K}_{input}. \tag{1}$$

Here, $\mathbf{K}_{task}$ is a $T \times T$ matrix describing the correlations between tasks, and $\mathbf{K}_{input}$ is a $N \times N$ matrix representing the input covariance.

A.2 Variational Inference

Employing a variational approach for inference in our MTGP model is pivotal for scalability and computational efficiency. Variational inference enables the approximation of the intractable posterior distribution of the GP with a variational distribution, mitigating the computational burdens typically associated with exact inference in GP models. This variational framework, complemented by the use of inducing points, allows for a lower-rank approximation to the GP prior, thereby rendering the model scalable to large datasets characteristic of urban air quality monitoring. We note that variational inference aims to approximate the true posterior distribution $p(\mathbf{f}|\mathbf{y})$ with a variational distribution $q(\mathbf{f})$:

$$KL[q(\mathbf{f})||p(\mathbf{f}|\mathbf{y})] = \int q(\mathbf{f}) \log \left(\frac{q(\mathbf{f})}{p(\mathbf{f}|\mathbf{y})} \right) d\mathbf{f}. \tag{2}$$

Optimizing this KL divergence yields the variational parameters defining $q(\mathbf{f})$.

Inducing Points: Inducing points $\mathbf{Z} = \{z_m\}_{m=1}^{M}$ facilitate a lower-rank approximation to the GP prior. The variational distribution over the inducing points $q(\mathbf{u})$ is typically chosen to be Gaussian:

$$q(\mathbf{u}) = \mathcal{N}(\mathbf{u}|\mathbf{m}, \mathbf{S}), \tag{3}$$

where $\mathbf{u} = \{f(z_m)\}_{m=1}^{M}$ represents the function values at the inducing points, and $\mathbf{m}$ and $\mathbf{S}$ are the mean vector and covariance matrix, respectively. The use of inducing points is a cornerstone for achieving scalability in our model. By forming a lower-rank approximation to the GP prior, inducing points facilitate a computationally efficient inference procedure without compromising the model's expressive power. The spatial distribution of inducing points, optimized during the variational inference procedure, further reflects the spatial heterogeneity inherent in the air quality data, providing insights into regions of higher information content.

A.3 Air-Quality Specific Adaptations

The Variational Spectral Multi-Task Gaussian Processes (VSMTGPs) model is formulated by combining the aforementioned components. The covariance function for VSMTGPs is given by:

$$\mathbf{K} = (\mathbf{K}_{task} \otimes \mathbf{K}_{input}) \otimes k_{SM}(x, x'), \tag{4}$$

where $\mathbf{K}_{task} \otimes \mathbf{K}_{input}$ encapsulates the multi-task and input space correlations, and $k_{SM}(x, x')$ models the temporal dependencies.

The variational lower bound on the log marginal likelihood is optimized to learn the model parameters, alongside the variational parameters of the inducing point distribution. This formulation enables a robust inference mechanism for air-quality modeling, capturing the spatial, temporal, and task correlations inherent in the data while ensuring computational scalability and efficiency.

The VSMTGP model is tailored to meet the unique challenges posed by air-quality inference. The model's capability to jointly learn the temporal and spatial correlations between different pollutants, along with its scalability afforded by the variational inference framework and inducing points, makes it a robust choice for modeling urban air quality dynamics. The elucidation of the variational distribution of inducing points provides a window into understanding the spatial distribution of information richness across the urban landscape, which is invaluable for policy-makers and urban planners aiming to enhance air quality monitoring and management.

Appendix B. Dataset

The dataset harnessed in this study encompasses hourly measurements of $PM_{2.5}$ and CO from 36 monitoring stations scattered across Beijing, supplemented by meteorological data from the respective district, spanning the period from May 1, 2014, to April 30, 2015. The meteorological data suite comprises temperature, humidity, pressure, wind speed, wind direction, and weather conditions, with wind direction and weather being categorized as categorical variables [6].

A substantial volume of data across varying stations and time intervals is missing, necessitating rigorous preprocessing. The preprocessing involved imputation techniques to fill in missing values, ensuring a consistent dataset for analysis. For the sake of comparative consistency with state-of-the-art neural baselines, the month of March 2015 was chosen due to its lower incidence of missing data.

Data Preprocessing An elaborate preprocessing routine was employed to address the issue of missing data, thereby ensuring data integrity for the ensuing analyses. Initial inspection revealed a significant absence of pressure data across the stations, which led to its exclusion from the analysis. Additionally, five stations (IDs: 1009, 1013, 1015, 1020, 1021) were discarded due to insufficient weather data.

The integrity of the remaining dataset was maintained by ensuring a minimum of 85% data availability for all variables, as detailed in Table 1. The missing data in real-valued variables ($PM_{2.5}$, temperature, humidity, and wind speed) were handled through time interpolation, a method chosen following a rigorous cross-validation exercise on non-missing data.