# *Focus on What's Important!* Inspecting Variational Distributions for Gaussian Processes for better *AQ* Station Deployment

**Progyan Das**[*]
Computer Science and Engineering
IIT Gandhinagar
India
`progyan.das@iitgn.ac.in`

**Mihir Agarwal**[*]
Electrical Engineering
IIT Gandhinagar
India
`agarwalmihir@iitgn.ac.in`

## Abstract

In urban locales, the intricate dynamics of air quality indicators such as Particulate Matter (PM2.5) and Carbon Monoxide (CO) necessitate sophisticated modeling for precise prediction and monitoring. However, monitoring stations are sparse, and effective placement is a key problem in the domain. This study explores a novel approach utilizing Variational Multi-Task Gaussian Processes (VMTGP) endowed with a Spectral Mixture (SM) kernel to model the spatiotemporal distribution of these pollutants in Beijing, which beats the state-of-the-art Gaussian Process techniques on this dataset in the exact MTGP case. However, our innovation lies in an in-depth examination of the variational distribution of the inducing points, which are critical for scalability and accurate approximations in GP models. Through an empirical lens, we observe a pronounced clustering of inducing points around certain monitoring stations, hinting at a higher information content in these locales. Our findings underscore the inherent value in exploiting the clustering phenomenon of inducing points, opening up new vistas for enhancing the efficacy and interpretability of multi-task learning paradigms in air quality forecasting. This insight holds promise for developing more robust and localized air quality prediction models, crucial for urban planning and public health policy formulations, and adaptively deciding the most effective locations for placing AQ monitoring stations.

Gaussian Process, Explainability, Adaptive Deployment, Air Quality

## 1 Introduction

Urban air quality monitoring is a critical endeavor that directly impacts public health and informs environmental policy [5] [6] [9]. The complex and interdependent nature of atmospheric pollutants necessitates sophisticated modeling techniques to accurately predict their behavior over time and space. Among the myriad of pollutants, Particulate Matter with a diameter of 2.5 micrometers or less (PM2.5) and Carbon Monoxide (CO) are primary indicators of air quality and possess significant health implications. PM2.5, due to its minute size, can infiltrate the respiratory system and engender severe health issues [11], while CO, a colorless and odorless gas, can be lethal when inhaled in large quantities [1]. The accurate prediction and monitoring of these pollutants are paramount for urban planning, public health policy formulation, and environmental conservation [4].

[*] These authors contributed equally to this research.

Computational Sustainability Workshop at NeurIPS 2023.

***Multi-Output Gaussian Processes*:** The spatiotemporal dynamics of PM2.5 and CO concentrations in urban settings like Beijing present a challenging prediction problem. The traditional single-output Gaussian Processes (SOGP) often fall short in capturing the underlying correlations between these pollutants. On the other hand, Multi-Output Gaussian Processes (MOGP) [3] offer a promising avenue to encapsulate the inherent correlations between multiple tasks, in this case, predicting the concentrations of PM2.5 and CO simultaneously. MOGPs, by design, allow for a shared representation of the input space, thus enabling the model to learn and exploit the temporal and spatial correlations between the tasks to improve prediction accuracy.

***Variational Distribution*:** A key aspect of our methodology is the analytical examination of the variational distribution of the inducing points [10], which are instrumental for scalability and accurate approximations in GP models. Our empirical findings reveal a distinct clustering of inducing points around certain monitoring stations, signifying higher information content in these locales. A meticulous validation demonstrates that these information-rich stations, when utilized individually for predictive tasks, exhibit a markedly lower Root Mean Square Error (RMSE) compared to their less-sampled counterparts.

## 2 Dataset and Preprocessing

The dataset leveraged in this investigation encapsulates hourly recordings of PM2.5 and CO from 36 surveillance stations dispersed throughout Beijing, augmented by meteorological data from the corresponding district, covering the duration from May 1, 2014, to April 30, 2015. The meteorological data ensemble encompasses temperature, humidity, pressure, wind speed, wind direction, and weather conditions, wherein wind direction and weather are designated as categorical variables. [2]

**Data Preprocessing** In addressing the challenge of missing data to ensure robust data integrity for subsequent analyses, an extensive preprocessing routine was deployed. Upon initial scrutiny, a marked lack of pressure data across various stations was observed, necessitating its omission from the analysis. Moreover, five stations, identified by IDs: 1009, 1013, 1015, 1020, 1021, were excluded due to the paucity of weather data. From these remaining stations, a random selection of four stations was made for experimental purposes, each bearing 8760 time-stamps over the month of March.

## 3 SVMT Spectral Gaussian Process

In addressing the challenges associated with urban air quality monitoring, a robust model capable of capturing the intricate dependencies between pollutants over time and space is imperative. Our Gaussian Process chosen model leans on a variational approach, employing a Linear Model of Coregionalization (LMC) [7], augmented with a Spectral Mixture (SM) kernel [8] and a Multitask kernel.

### 3.1 Variational Approach

Variational methods provide a mechanism to approximate intractable posterior distributions by optimizing a variational distribution to minimize the divergence from the true posterior. Mathematically, given the observed data $\mathbf{y}$, we aim to approximate the posterior $p(\mathbf{f}|\mathbf{y})$ by a variational distribution $q(\mathbf{f})$, where $\mathbf{f}$ denotes the latent functions. The optimization objective is to minimize the Kullback-Leibler divergence between $q(\mathbf{f})$ and $p(\mathbf{f}|\mathbf{y})$, given as:

$$KL[q(\mathbf{f})||p(\mathbf{f}|\mathbf{y})] = \int q(\mathbf{f}) \log \left( \frac{q(\mathbf{f})}{p(\mathbf{f}|\mathbf{y})} \right) d\mathbf{f}. \tag{1}$$

Inducing points, denoted $\mathbf{Z} = \{z_m\}_{m=1}^{M}$, where $M$ is the number of inducing points, are introduced to form a lower-rank approximation to the GP prior. The variational distribution over the inducing points $q(\mathbf{u})$ is optimized, where $\mathbf{u} = \{f(z_m)\}_{m=1}^{M}$ represents the function values at the inducing points. Typically, $q(\mathbf{u})$ is chosen to be a multivariate Gaussian distribution with a mean vector $\mathbf{m}$ and a covariance matrix $\mathbf{S}$: $q(\mathbf{u}) = \mathcal{N}(\mathbf{u}|\mathbf{m}, \mathbf{S})$.

The variational parameters $\mathbf{m}$ and $\mathbf{S}$ are optimized to minimize the KL divergence, and the inducing points $\mathbf{Z}$ are either fixed a priori or optimized jointly with the variational parameters.

### 3.2 Linear Model of Coregionalization (LMC)

LMC facilitates a joint analysis that leverages the correlations between these pollutants. LMC posits that each task is a linear combination of shared latent functions. Let $f_d(x)$ denote the latent functions for $D$ outputs. The LMC model can be expressed as: $g_k(x) = \sum_{d=1}^{D} B_{kd} f_d(x)$, where $g_k(x)$ is the $k$-th output function, and $B$ is a coregionalization matrix.

### 3.3 Spectral Mixture (SM) Kernel and Multitask Kernel

The SM kernel is instrumental in modeling the complex periodic patterns often inherent in air quality data. The SM kernel is expressed as a sum of $Q$ Gaussian mixtures: $k_{SM}(x, x') = \sum_{q=1}^{Q} w_q \exp\left(-2\pi^2\tau_q^2\|x - x'\|^2\right)\cos\left(2\pi\mu_q(x - x')\right)$, where $w_q, \tau_q$, and $\mu_q$ are the weights, lengthscales, and means of the mixtures, respectively.

The Multitask kernel allows for the modeling of correlations between tasks, extending the SM kernel to a multi-task scenario. The Multitask kernel is expressed as a Kronecker product between the SM kernel and a task covariance matrix $K_{task}$: $k(x, x', k, k') = k_{SM}(x, x') \otimes K_{task}(k, k')$.

### 3.4 Evaluation and Results

| Model | $PM_{2.5}$ | CO |
|---|---|---|
| **Multi-Output Exact GP** | **26.79** | **0.46** |
| Single Output Exact GP | 27.67 | 0.63 |
| Multi-Output Sparse GP (1000 Inducing Points) | 45.76 | 0.71 |
| Single Output Sparse GP (1000 Inducing Points) | 47.27 | 0.81 |
| Multi-Layer Perceptron | 79.56 | 0.74 |

Table 1: RMSE for multitask forecasting over (latitude, longitude, temperature, humidity, wind speed)

On both $PM_{2.5}$ and CO, our Multi-Output Exact GP with LMC and SM Kernel beats the other state-of-the-art Single-Output Exact GP, which has been shown to beat neural-attention baselines with the right choice of kernels. We leverage this and use the same MOGP with a variational strategy for selecting inducing points to make sure we get a robust variational distribution that represents our data very well.

### 3.5 Inspection of Variational Distribution

We inspect the variational distribution by considering the inducing points as a sampling from this distribution. By plotting the latitude and longitude of these inducing points alongside the stations, we observe the spatial relationship between the inducing points and the monitoring stations.

We employed the Euclidean distance to associate each inducing point with the nearest monitoring station, thereby creating a mapping that reflects which stations capture which inducing points. For longer distances, the Haversine distance, which takes into account the Earth's curvature, would provide a more accurate measure. We observe that certain stations are more adept at "capturing" inducing points, indicative of these stations being proximal to regions where the variational distribution peaks. This observation infers a measure of the *information richness* of different geographical locales with respect to the modeled pollutants. The peaking of the variational distribution around particular locations accentuates the regions of higher information content, which, in turn, highlights the critical stations that significantly contribute to the accuracy and interpretability of the predictive model.

## 4 Discussion

The analysis of the variational distribution of inducing points in our Variational Spectral Multi-Task Gaussian Processes model has unveiled regions of higher information richness, predominantly around certain air quality monitoring stations. These regions, by virtue of being information-rich, provide a robust basis for the model to learn and make accurate predictions concerning air quality indicators. However, this also casts light on the concomitant regions that are less well-explored, and hence, characterized by lower information richness.

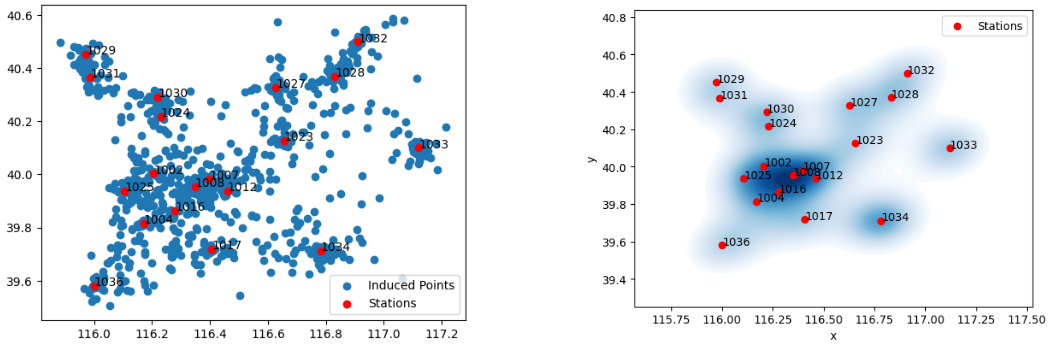

Figure 1: *Left*: The latitude/longitude indices of the inducing points sampled from the variational distribution plotted with the location of the AQ monitors. *Right*: A Kernel Distribution Estimation (KDE) plot of the inducing points, as an approximation of the variational distribution

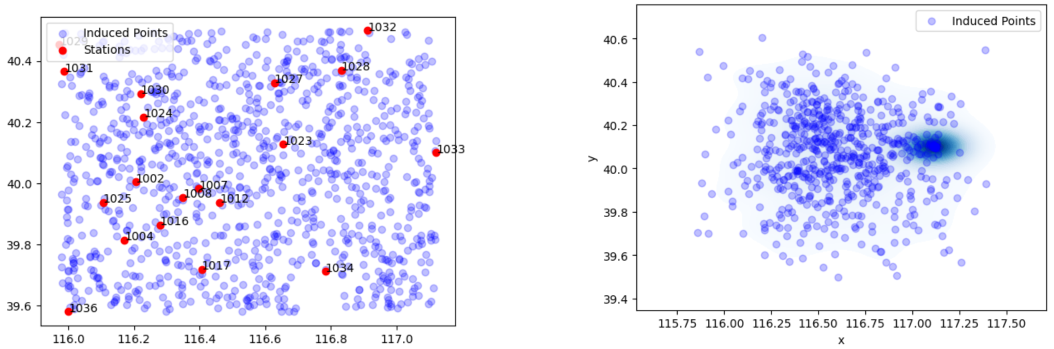

Figure 2: *Left*: The inducing point sampling for an untrained Gaussian Process. *Right*: KDE plot when the Gaussian process has been only trained on one station – the distribution peaks at that station.

Actively deploying additional air quality (AQ) monitoring stations in these less well-explored zones is a strategic maneuver to bridge the information gap. This deployment could substantially enhance the spatial coverage and granularity of the air quality monitoring network, thereby ameliorating the model's capability to capture the spatial heterogeneities in air pollutant concentrations.

## 4.1 Implications for Urban Planning and Public Health

From an urban planning and public health perspective, having a finely grained understanding of air quality across the urban landscape is indispensable. The additional data collected from the newly deployed AQ stations in the less well-explored zones would not only enrich the dataset but also potentially unveil localized pollution sources or pollution hotspots.

**Enhanced Predictive Performance** The infusion of data from the less well-explored zones is likely to enhance the predictive performance of our model. A more balanced representation of the urban environment, encompassing both information-rich and information-poor regions, would furnish the model with a holistic understanding of the spatial dynamics governing air pollutant concentrations.

**Adaptive Monitoring Strategies** This study also lays the groundwork for the development of adaptive monitoring strategies. Adaptive deployment or re-deployment of AQ monitoring stations in response to such shifts could ensure that the monitoring network remains optimally configured to capture the evolving air quality dynamics.

In conclusion, the active deployment of AQ monitoring stations in less well-explored zones is a prudent step towards achieving a comprehensive and nuanced understanding of urban air quality. This understanding is instrumental for informed urban planning and policy formulation aimed at ameliorating air quality and safeguarding public health.

## 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] 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.

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

[4] 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.

[5] 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.

[6] 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.

[7] Pablo Moreno-Muñoz, Antonio Artés, and Mauricio Alvarez. "Heterogeneous multi-output Gaussian process prediction". In: *Advances in neural information processing systems* 31 (2018).

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

[9] 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.

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

[11] 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.