POLITECNICO
MILANO 1863

Effect of Area C Introduction

Expansion on PM10 and PM2.5 concentrations

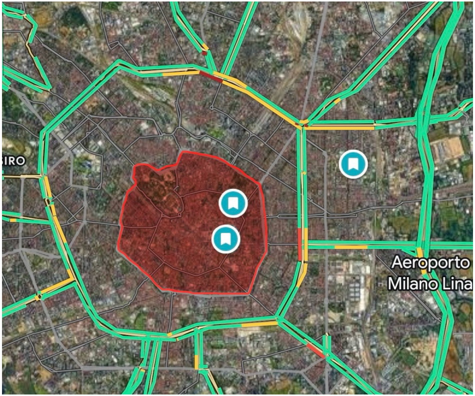
Students: Gutierrez M., Luca A., Maiocchi E., Mazza M., Sottanella A., Travaglini F.

Tutors: Colombara S., Beltramin G.

Dataset Introduction

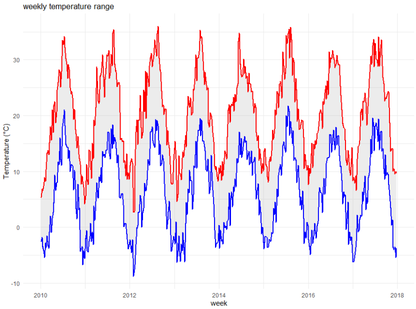
Our data comes from the union of two different datasets:

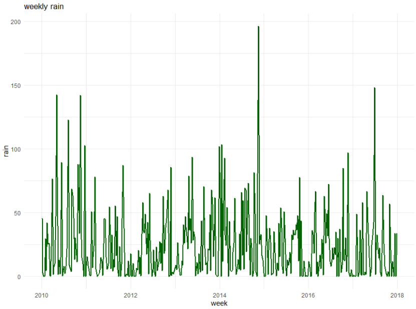
- The daily measurements of pollutants (**PM10** and **PM2.5**) collected from 2010 to 2017 by multiple **ARPA** monitoring stations located in various areas of Milan;
- The daily recordings of **weather conditions** from the website **OpenMeteo**.

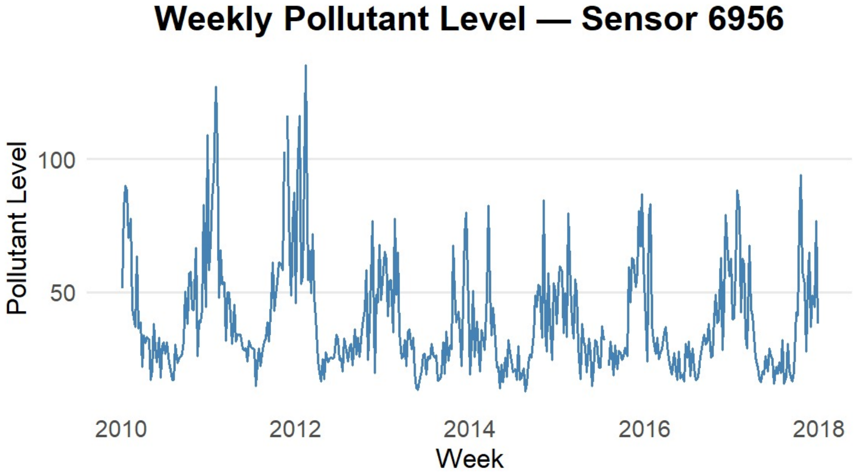


Exploratory Analysis

We took the weekly average from the pollutants and used other aggregation methods for each weather condition.

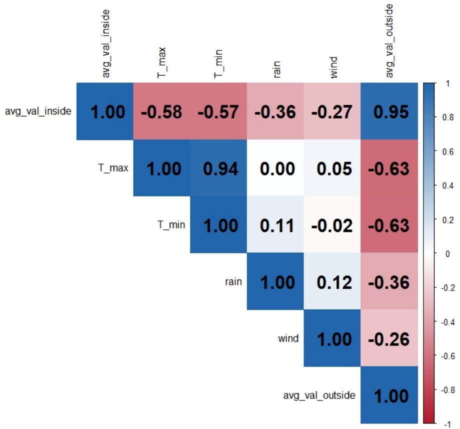





Weekly Pollutant Level — Sensor 6956
Pollutant Level
100
50
2010
2012
2014
2016
2018
Week

Exploratory Analysis

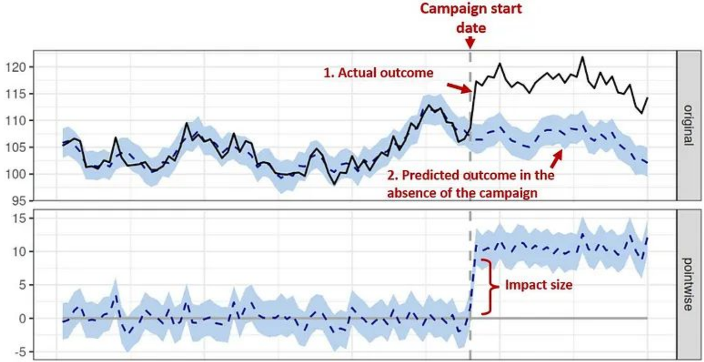
Here we have the correlation plot with the weather conditions and the two PM10 sensors inside and outside of area C.



Causal Impact: general idea

Goal: estimate the effect of an intervention on an outcome of interest by comparing what actually happened to an estimate of what would have happened in the absence of the intervention. *[Brodersen et al. (2015)]*

Our case: the intervention is the introduction of **Area C** (Jan 2012) in Milan.



Causal Impact: Our Model

$$y_t \mid \mu_t, \gamma_t, \boldsymbol{\beta}, \sigma_\varepsilon^2 \overset{\text{ind}}{\sim} \mathcal{N}\left(\mu_t + \gamma_t + \mathbf{x}_t^\top \boldsymbol{\beta},\ \sigma_\varepsilon^2\right) \qquad t = 1, \ldots, n$$

where:

- y_t is the average PM10 concentration inside Area C;
- μ_t is the local linear level (baseline trend);
- γ_t is the seasonal component;
- $\mathbf{x}_t$ contains the covariates values (weather and control sensor);
- $\boldsymbol{\beta}$ contains the regression coefficients;
- σ_ε^2 is the observation noise variance.

Causal Impact: Our Model

$$y_t \mid \mu_t, \gamma_t, \boldsymbol{\beta}, \sigma_\varepsilon^2 \overset{\text{ind}}{\sim} \mathcal{N}\left(\mu_t + \gamma_t + \mathbf{x}_t^\top \boldsymbol{\beta},\ \sigma_\varepsilon^2\right) \qquad t = 1, \ldots, n$$

- **Baseline Trend:**

$$\mu_{t+1} = \mu_t + \eta_{\mu,t}, \quad \eta_{\mu,t} \sim N(0, \sigma_\mu^2)$$

- **Seasonality:**

$$\gamma_{t+1} = -\sum_{s=0}^{50} \gamma_{t-s} + \eta_{\gamma,t}, \qquad \sum_{s=0}^{51} \gamma_{t-s} = 0, \qquad \eta_{\gamma,t} \sim \mathcal{N}(0, \sigma_\gamma^2)$$

- **Variances:**

$$\frac{1}{\sigma_\mu^2}, \frac{1}{\sigma_\gamma^2}, \frac{1}{\sigma_\varepsilon^2} \overset{\text{ind}}{\sim} \text{Gamma}\left(\frac{\nu_i}{2}, \frac{s_i}{2}\right), i = \mu, \gamma, \epsilon$$

Causal Impact: Our Model

$$y_t \mid \mu_t, \gamma_t, \boldsymbol{\beta}, \sigma_\varepsilon^2 \overset{\text{ind}}{\sim} \mathcal{N}\left(\mu_t + \gamma_t + \mathbf{x}_t^\top \boldsymbol{\beta},\ \sigma_\varepsilon^2\right) \qquad t = 1, \dots, n$$

- **Regression Component**

$$\mathbf{x}_t^\top \boldsymbol{\beta} = \beta_1 T_{\text{max},t} + \beta_2 T_{\text{min},t} + \beta_3 \,\text{rain}_t + \beta_4 \,\text{wind}_t + \beta_5 \, y_t^{\text{outside}}$$

- **Spike-and-slab prior**

$$\beta_j \mid \gamma_j \overset{\text{ind}}{\sim} (1 - \gamma_j)\,\delta_0 + \gamma_j \,\mathcal{N}\left(0, \sigma_{\beta_j}^2\right) \qquad j = 1, \dots, 5$$

$$\gamma_j \mid \theta_j \overset{\text{ind}}{\sim} \text{Be}(\theta_j)$$

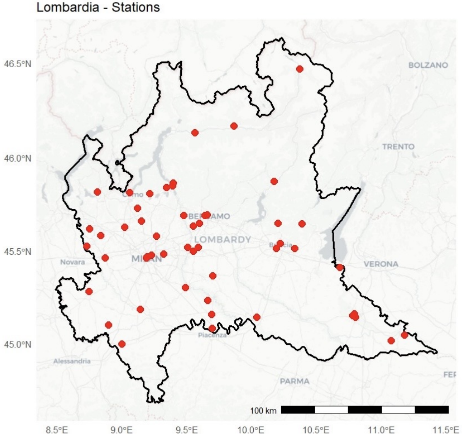
$$\theta_j \overset{\text{ind}}{\sim} \pi(\theta_j)$$

Causal Impact: Results

From the cumulative difference plot we can see a reduction of PM10 after the introduction of Area C.
We obtained a posterior probability of an effect equal to 97.236%.

Spatial Stan Model For Lombardy

Lombardy has 99 ARPA monitoring stations for PM10 in total but only **51 stations** actively and continuously recorded data during the study period from the beginning of 2010 to the end of 2017.



Spatial Stan Model For Lombardy

$$\mathbf{y}_t | \alpha, \mathbf{y}_{t-1}, \boldsymbol{\beta}, \mathbf{w}, \Sigma \overset{\text{ind}}{\sim} \mathcal{N}_M\big(\alpha \mathbf{y}_{t-1} + \boldsymbol{\beta}\mathbf{X}_t + \mathbf{w}, \Sigma\big) \qquad t = 1, \ldots, n$$

where:

- $\mathbf{y}_t \in \mathbb{R}^M$ is the vector of pollutant concentrations observed at time t across the $M = 51$ stations;
- $\mathbf{X}_t$ contains the $K = 4$ covariates at time t (max and min temperature, sum of the weekly rain and max wind speed);
- $\alpha \in (-1, 1)$ is the autoregression coefficient;
- $\boldsymbol{\beta} \in \mathbb{R}^K$ is the vector of regression coefficients;
- $\mathbf{w} \in \mathbb{R}^M$ models spatial random effects.

Spatial Stan Model For Lombardy

$$\mathbf{y}_t | \alpha, \mathbf{y}_{t-1}, \boldsymbol{\beta}, \mathbf{w}, \Sigma \overset{\text{ind}}{\sim} \mathcal{N}_M\left(\alpha \mathbf{y}_{t-1} + \boldsymbol{\beta} \mathbf{X}_t + \mathbf{w}, \Sigma\right) \qquad t = 1, \ldots, n$$

- **Spatial effect:**

$$\mathbf{w} \sim \mathcal{N}_M(\mathbf{0}, S_w),$$

$$S_w(i,j) = \sigma_w^2 \exp\left(-\frac{\text{dist}(s_i, s_j)}{L}\right), \qquad i, j = 1, \ldots, M$$

$$\sigma_w^2 \sim Inverse - Gamma(2, 1),$$

where s_i and s_j denote the location of stations i and j
while $L = max(\frac{\text{dist}(s_i, s_j)}{3})$

Spatial Stan Model For Lombardy

$$\mathbf{y}_t | \alpha, \mathbf{y}_{t-1}, \boldsymbol{\beta}, \mathbf{w}, \Sigma \stackrel{\text{ind}}{\sim} \mathcal{N}_M(\alpha \mathbf{y}_{t-1} + \boldsymbol{\beta} \mathbf{X}_t + \mathbf{w}, \Sigma) \qquad t = 1, \dots, n$$

- **Autoregressive component:**

$$\alpha \sim \text{Uniform}(-1, 1)$$

- **Regression coefficients:**

$$\boldsymbol{\beta} \sim \mathcal{N}_K(\mathbf{0}, I) \qquad K = 1, \dots, 4$$

- **Covariance:**

$$\Sigma \sim Inverse - Wishart(M + 2, \Sigma_0),$$

where Σ_0 is the empirical covariance matrix.

This was computationally infeasible!

$$\mathbf{y}_t | \alpha, \mathbf{y}_{t-1}, \boldsymbol{\beta}, \mathbf{w}, \Sigma \overset{\text{ind}}{\sim} \mathcal{N}_M(\alpha \mathbf{y}_{t-1} + \boldsymbol{\beta}\mathbf{X}_t + \mathbf{w}, \Sigma) \qquad t = 1, \dots, n$$

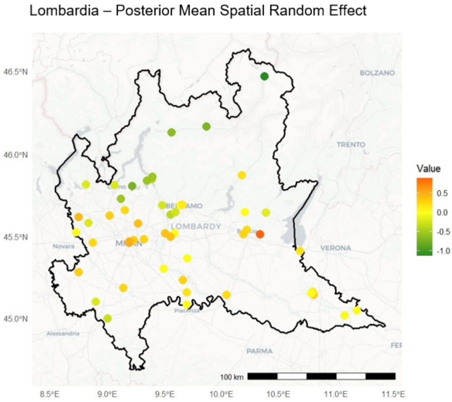
we decomposed Σ as

$$\Sigma = diag(\tau) \times \Omega \times diag(\tau),$$

where $diag(\tau)$ is a diagonal matrix of standard deviations and Ω is a correlation matrix. Ω was modeled using an **LKJ distribution**. Then, we imposed additional sparsity by constraining to zero the off-diagonal elements of Ω corresponding to pairs of stations separated by more than **100 km**.

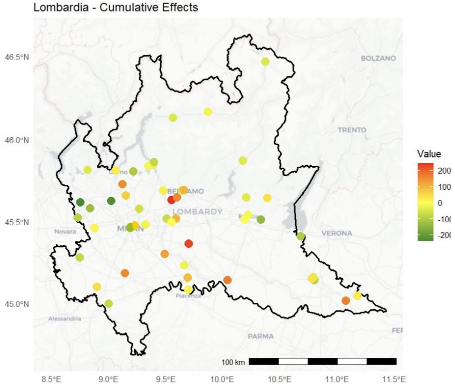
Spatial Stan Model For Lombardy: Results

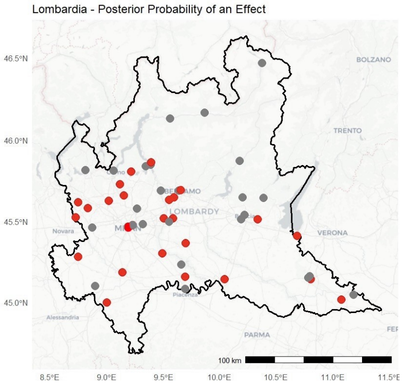
- $\alpha \in [0.42, 0.47]$: strong temporal persistence in pollutant concentration;
- The regression coefficients $\boldsymbol{\beta}$ all have credible intervals that include zero;
- The spatial random effects $\mathbf{w}$ reveal a clear geographical pattern.



Spatial Stan Model For Lombardy: Results

Cumulative effect (on the left) and Posterior probability of an effect (on the right: red dots for stations with strong evidence) for each station.





References

- Brodersen, Kay H., et al. (2015). Inferring causal impact using Bayesian structural time-series models. Annals of Applied Statistics: 247–274.
- http://google.github.io/CausalImpact/CausalImpact.html.
- Stan User's Guide. Multivariate priors for hierarchical models. https://mc-stan.org/docs/stan-users-guide/regression.html
- https://github.com/marymazza/BayesianStatisticsProject