1921 — 2021 UNIVERSITÀ CATTOLICA del Sacro Cuore

Statistical and actuarial sciences

Data analytics for business

RESTful scraping API for Real Estate Rental data, a Bayesian Spatial modeling perspective with INLA

Niccolò Salvini ID: 4806876
Candidate

PhD Marco L. Della Vedova
Supervisor

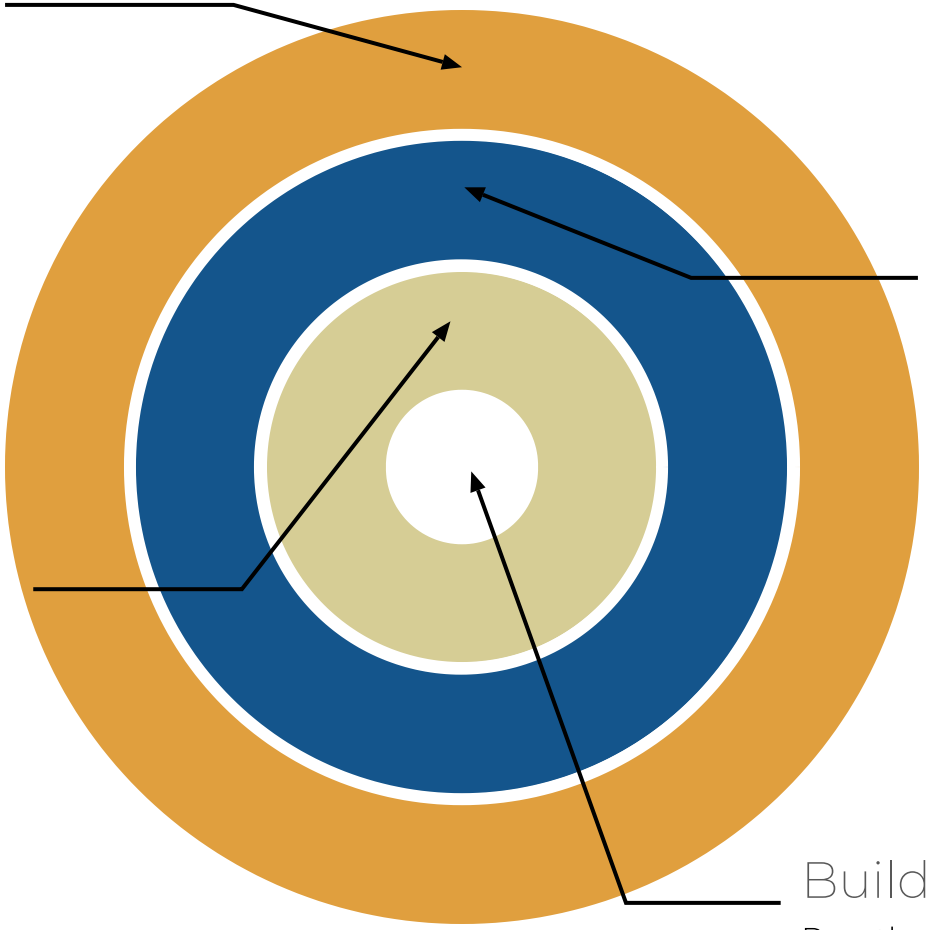
PhD Vincenzo Nardelli
Assistant Supervisor

A.Y. 2020 - 2021

Research Question

Obtain **Data**

Extract Rental data



Build the **Software**

A way to sequentially and quickly extract data

Adjust settings to use the **Algorithm**

Data might be massive, we need an algorithm to reduce computational time

Build a **Model** and **Results**

Run the model & evaluate results

The following work has the aim to build a robust **Scraping API** service to extract **Real Estate rental Data** (Milan, IT) and applying **Geostatistics spatial** modeling through a convenient computing alternative called **INLA**.

Agenda

Section 2

Web Scraping

How to extract data

INLA

A faster estimation algorithm

Start — *End*

RESTful API

The web service that makes it possible

Spatial Model on Real Estate (rental)

Application and Results on data scraped

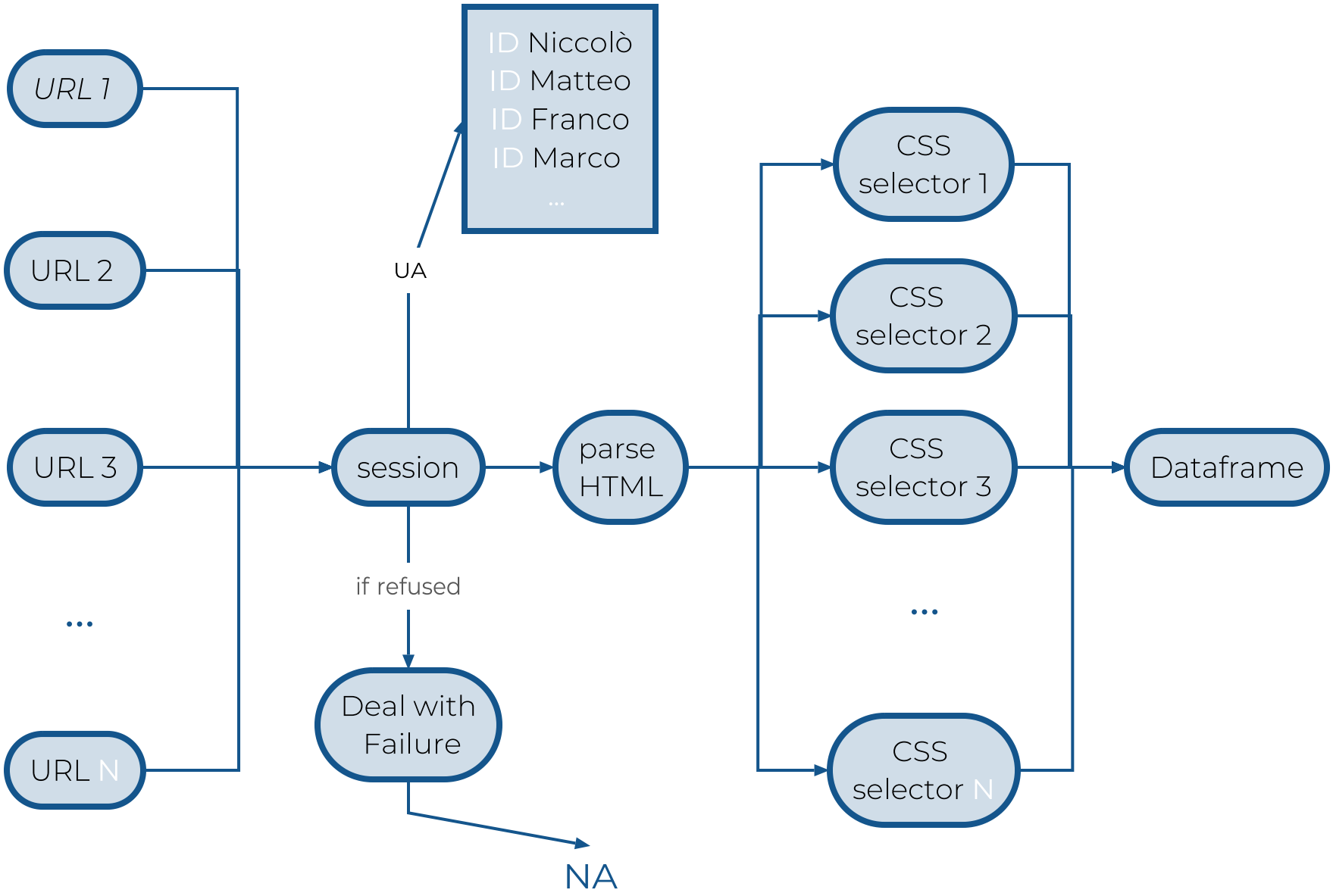
Section 1

Section 1.1

Web Scraping

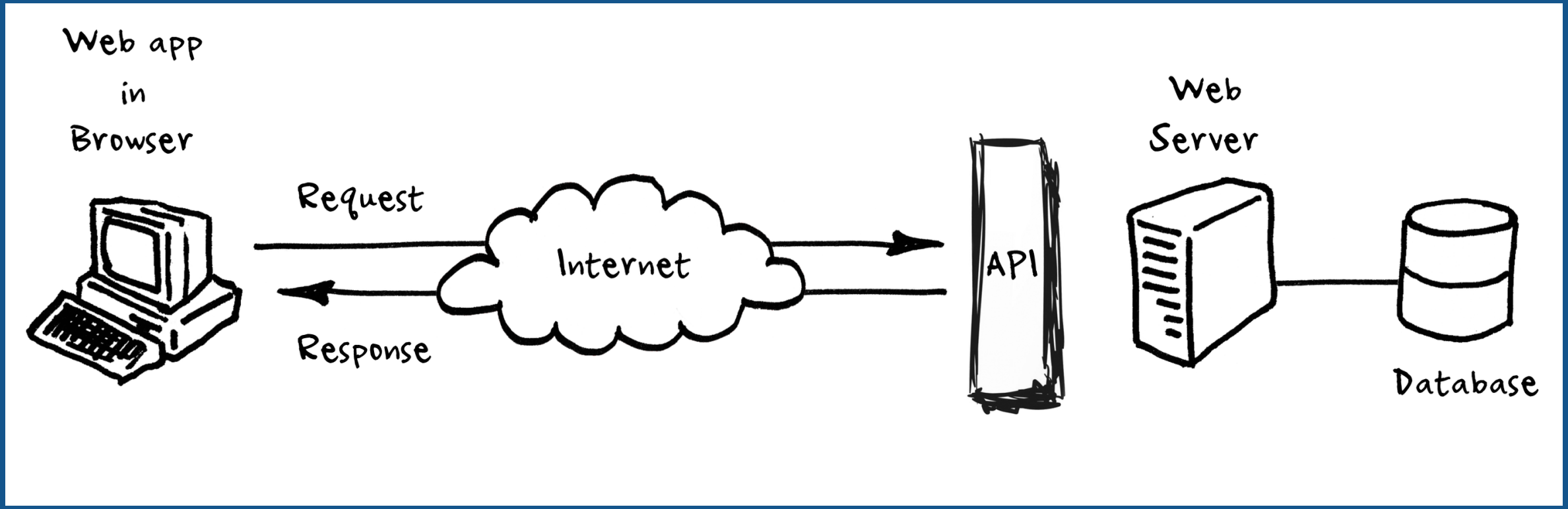
Web Scraping

Web Scraping is a technique aimed at **extracting unstructured** data from static or dynamic internet **web pages** and **collecting** it in a **structured** way.



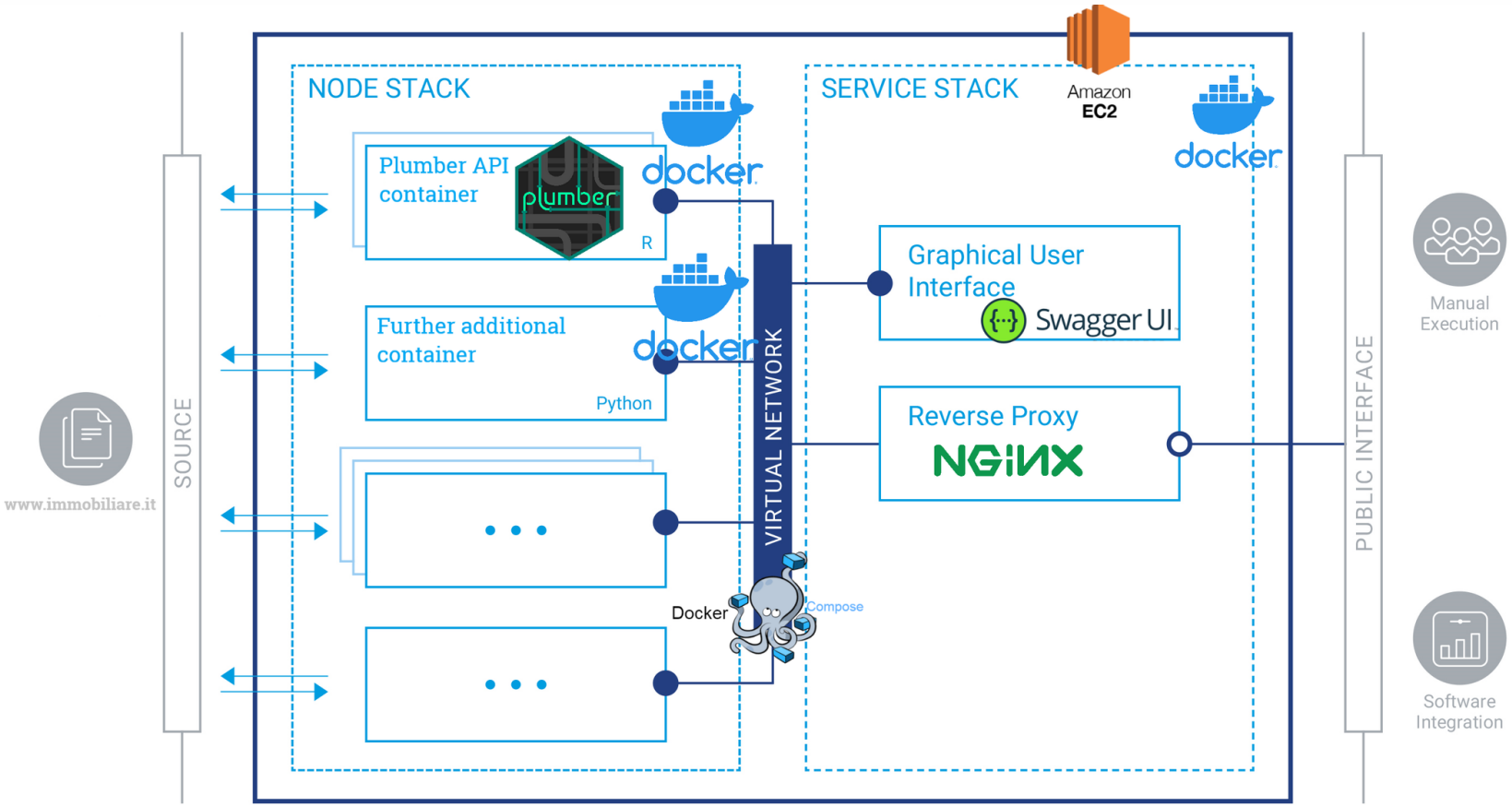
Section 1.1

RESTful API



RESTful API

APIs let a product or a service **communicate** with other products and services **without** having to **know how** they're **implemented**.



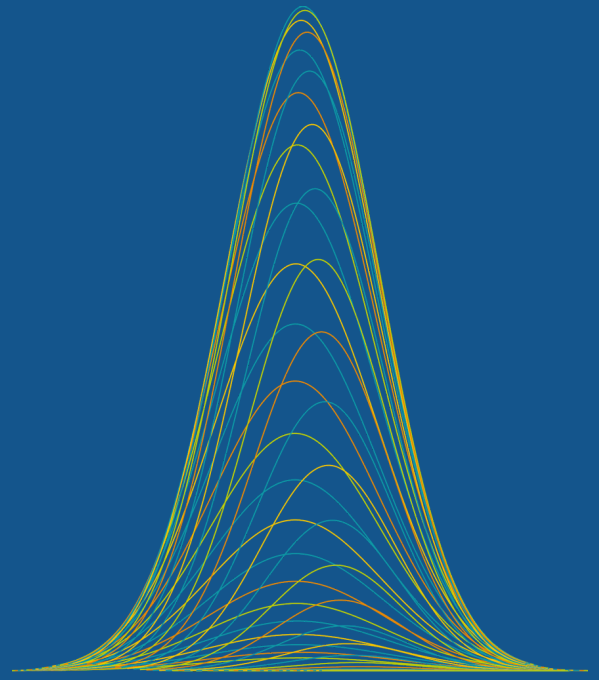
API call results

Data depend on the API arguments specified, covariates scraped are **64**, including **Latitude** and **Longitude.** Here a glance of the first 4 rows per 9 columns.

ID	Latitude	Longitude	Location	Condomini	Building Age	Rooms	AC	Monthly Price	…
84095940	45.460	9.189	via paolo da cannobio 37	€ 200/mese	1950	2 (1 camera da letto, 1 altro), 1 bagno, cucina angolo cottura	Autonomo, freddo	€ 1.350/mese	…
82824303	45.484	9.171	via tartaglia 7	€ 250/mese	1965	2 (1 camera da letto, 1 altro), 1 bagno, cucina semi abitabile	Autonomo, freddo/caldo	€ 1.200/mese	…
82693539	45.470	9.208	via antonio kramer 20	€ 133/mese	1920	3 (2 camere da letto, 1 altro), 1 bagno, cucina a vista	Autonomo, freddo/caldo	€ 1.600/mese	…
83486379	45.474	9.186	via solferino 11	€ 175/mese	2003	5 (3 camere da letto, 2 altri), 2 bagni, cucina abitabile	Autonomo, freddo/caldo	€ 3.250/mese	…
…	…	…	…	…	…	…	…	…	

Section 2.1

INLA algorithm



INLA

INLA Algorithm

Integrated Nested Laplace Approximation constitutes a faster and accurate deterministic algorithm based on a special type of models called Latent Gaussian Models, **LGM.**

INLA turns out to **shorten model fitting time** for essentially **2 reasons**: Gaussian Markov random field (**GMRF**) offering sparse matrices representation and **Laplace Approximation** to approximate posterior marginals' integrals with proper search strategies.

Since parameters and Hyper parameters are many a **Hierarchical Structure** is imposed

Notation Remark: *Bold symbols are vectors*

$$\boldsymbol{\theta} = \{\beta_0, \boldsymbol{\beta}, f\} \qquad \boldsymbol{\psi} = \{\boldsymbol{\psi}_1, \boldsymbol{\psi}_2\}$$

Linear Predictor

$$\eta_i = \beta_0 + \sum_{m=1}^{M} \beta_m \boldsymbol{x}_{mi} + \sum_{l=1}^{L} \boldsymbol{f}_l(z_{li})$$

- To start it can be specified a **generalization** of a **linear predictor η** which takes into account both **linear** and **non-linear effects** on covariates.
- **Non-linear** effects on covariates are really **important** since it is the part **where** the statistical model **integrates** the **spatial component** in the form of a *Gaussian Process*.
- Imagine to group all the parameters into a vector said *latent field* ***theta***.

INLA Algorithm

Integrated Nested Laplace Approximation constitutes a faster and accurate deterministic algorithm based on a special type of models called Latent Gaussian Models, **LGM.**

INLA turns out to **shorten model fitting time** for essentially **2 reasons**: Gaussian Markov random field (**GMRF**) offering sparse matrices representation and **Laplace Approximation** to approximate posterior marginals' integrals with proper search strategies.

Since parameters and Hyper parameters are many a **Hierarchical Structure** is imposed

Notation Remark: *Bold symbols are vectors*

$$\boldsymbol{\theta} = \{\beta_0, \boldsymbol{\beta}, f\} \qquad \boldsymbol{\psi} = \{\boldsymbol{\psi}_1, \boldsymbol{\psi}_2\}$$

$$\pi(\boldsymbol{\theta}, \boldsymbol{\psi} \mid \mathbf{y}) \propto \underbrace{\pi(\boldsymbol{\psi})}_{\text{priors}} \times \underbrace{\pi(\boldsymbol{\theta} \mid \boldsymbol{\psi})}_{\text{GMRF}} \times \underbrace{\prod_{i=1}^{\mathbf{I}} \pi(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\psi})}_{\text{likelihood}}$$

Latent Gaussian Models

- They are composed by three nested hierarchy levels:
- the *higher* **Likelihood** of data
- the *medium* **Latent Gaussian Random Field** (also said latent field, where parameters are, for which priors have to be specified
- the *lower* ***Priors***

INLA Algorithm

Integrated Nested Laplace Approximation constitutes a faster and accurate deterministic algorithm based on a special type of models called Latent Gaussian Models, **LGM.**
INLA turns out to **shorten model fitting time** for essentially **2 reasons**: Gaussian Markov random field (**GMRF**) offering sparse matrices representation and **Laplace Approximation** to approximate posterior marginals' integrals with proper search strategies.
Since parameters and Hyper parameters are many a **Hierarchical Structure** is imposed

Notation Remark: *Bold symbols are vectors*

$$\boldsymbol{\theta} = \{\beta_0, \boldsymbol{\beta}, f\} \qquad \boldsymbol{\psi} = \{\boldsymbol{\psi}_1, \boldsymbol{\psi}_2\}$$

$$\pi(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\psi_1}) = \prod_{i=1}^{\mathbf{I}} \pi(y_i \mid \theta_i, \boldsymbol{\psi_1})$$

Likelihood < *higher* >

- Likelihood of data, generally are chosen **Exponential Family distribution** function, i.e. Normal, Poisson etc.
- **Theta** is the latent field
- **Psi_1** are hyper parameters retraceable to the first level of the Hierarchy.

INLA Algorithm

Integrated Nested Laplace Approximation constitutes a faster and accurate deterministic algorithm based on a special type of models called Latent Gaussian Models, **LGM.**

INLA turns out to **shorten model fitting time** for essentially **2 reasons**: Gaussian Markov random field (**GMRF**) offering sparse matrices representation and **Laplace Approximation** to approximate posterior marginals' integrals with proper search strategies.

Since parameters and Hyper parameters are many a **Hierarchical Structure** is imposed

Notation Remark: *Bold symbols are vectors*

$$\boldsymbol{\theta} = \{\beta_0, \boldsymbol{\beta}, f\} \qquad \boldsymbol{\psi} = \{\boldsymbol{\psi}_1, \boldsymbol{\psi}_2\}$$

GMRF < *medium* >

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{\psi}_2) = \mathrm{MVN}(0, \boldsymbol{Q}(\boldsymbol{\psi}_2))$$

- it is just a random vector distributed as a Multivariate Normal distribution with mean **0** and precision matrix **Q(psi_2)** (priors coming from the medium level Matérn covariance).
- It takes the name of **Markov** since it enjoys special **properties** encoded in the precision matrix due to a conditional independence assumption making it **sparse.**
- Gaussian priors must be taken

INLA Algorithm

Integrated Nested Laplace Approximation constitutes a faster and accurate deterministic algorithm based on a special type of models called Latent Gaussian Models, **LGM.**
INLA turns out to **shorten model fitting time** for essentially **2 reasons**: Gaussian Markov random field (**GMRF**) offering sparse matrices representation and **Laplace Approximation** to approximate posterior marginals' integrals with proper search strategies.
Since parameters and Hyper parameters are many a **Hierarchical Structure** is imposed

Notation Remark: *Bold symbols are vectors*

$$\boldsymbol{\theta} = \{\beta_0, \boldsymbol{\beta}, f\} \qquad \boldsymbol{\psi} = \{\boldsymbol{\psi}_1, \boldsymbol{\psi}_2\}$$

Priors < *lower* >

$$\pi(\boldsymbol{\psi}), \quad \boldsymbol{\psi}\{\boldsymbol{\psi_1}, \boldsymbol{\psi_2}\}$$

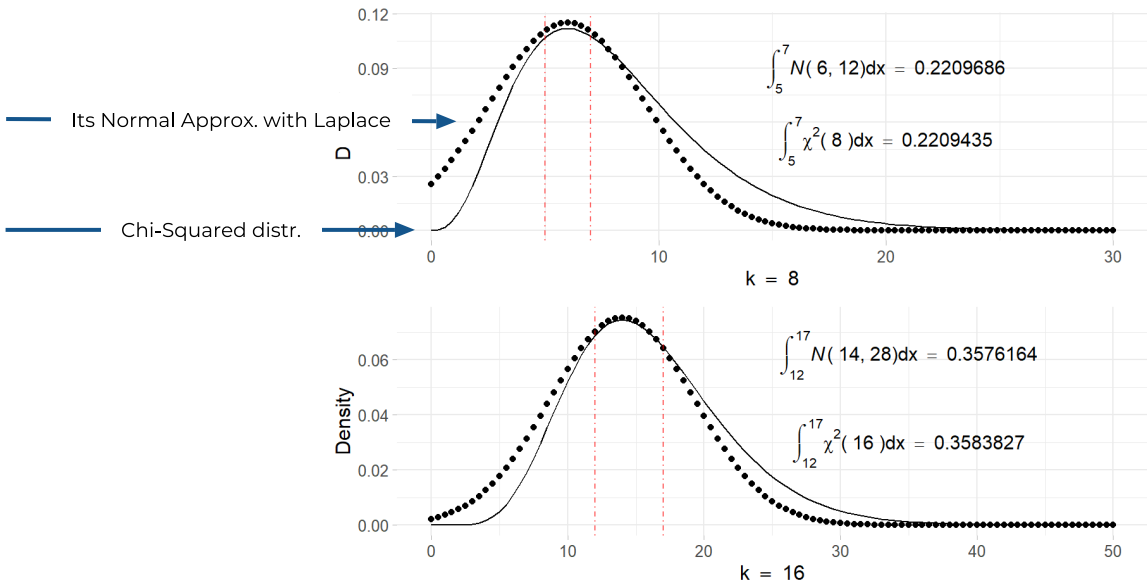
- Priors need to be specified also for hyper parameters as their joint probability distribution or their distinct product.
- No distribution constraints

Laplace Approx

Bayesian inference is interested into **measure posterior** and other quantities of interest for marginal distribution for **each element** in the **latent field** and the **hyper-parameter** vector. One way to compute this quantities is via ***Approximation*** (w.r.t. *Sampling* or *Exact*, i.e. conjugate distr.). **Tilde** are the Laplace approximation for the correspondent distribution.

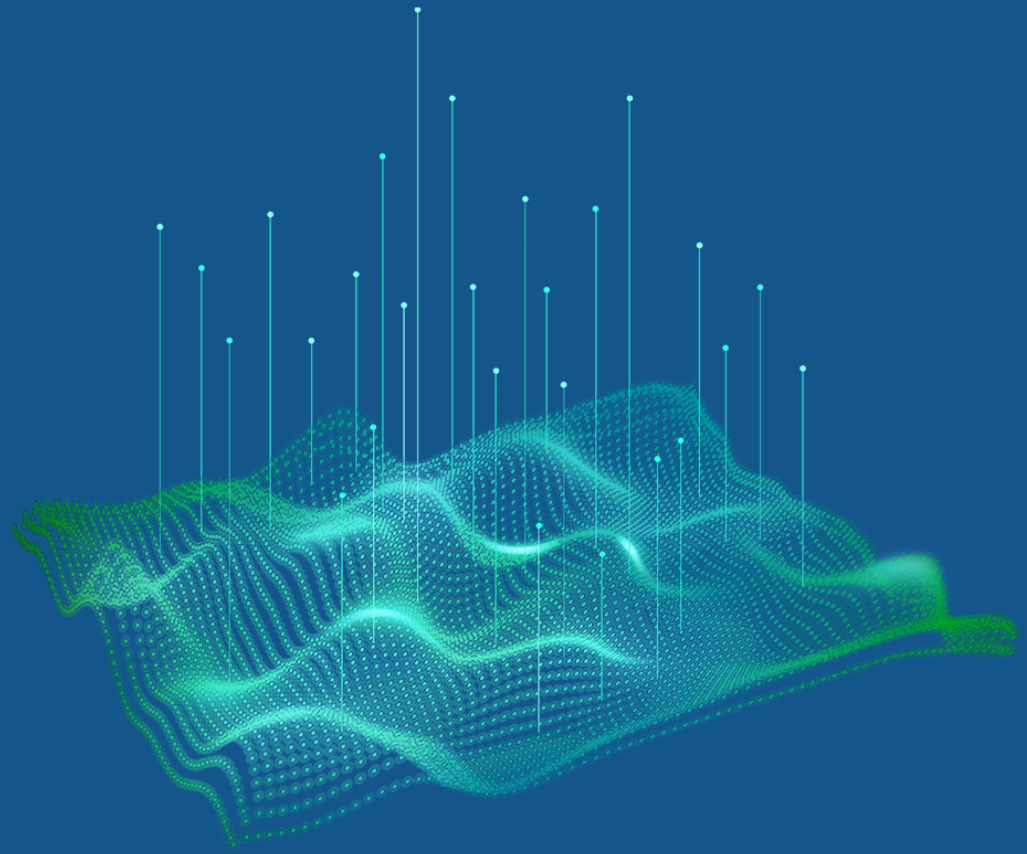
$$\tilde{\pi}(\theta_i \mid \mathbf{y}) = \int \pi(\theta_i \mid \psi, \mathbf{y}) \pi(\psi \mid \mathbf{y}) d\psi$$

$$\tilde{\pi}(\psi_k \mid \mathbf{y}) = \int \pi(\psi \mid \mathbf{y}) d\psi_{-k}$$



Section 2.2

Model Fitting & Results



Fit INLA

Data come from **Milan** Real Estate Rental within the municipality borders, extracted in **date: 22nd January 2021**

$$g(\mu_i) = \eta_i$$

$$\xi_i \sim N\left(\mathbf{0}, \mathbf{Q}_{\mathscr{C}}^{-1}\right)$$

$$\boldsymbol{\theta} = \{\boldsymbol{\xi}, \boldsymbol{\beta}\}$$

$$\boldsymbol{\eta} = b_0 + \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{\xi}$$

$$\psi_1 = \left\{\sigma_\varepsilon^2\right\}$$

The model Error

$$\psi_2 = \left(\sigma_\xi^2, \phi, \nu\right)$$

Variance GP

Hyper Params from Matèrn

Linear Predictor

- **x** are all the covariates scraped (recall slide 8) and **betas** are the respective coefficients, assigned Gaussian Priors
- The function **g()** is identity, linear mapping from mean to **Eta**
- Recall generlized lienar predictor, f is Gaussian Process **XI** t for the **Z** set of covariates Lat and Long for which a **GMRF** is specified.
- **beta_0** is the intercept.
- The theoretical foundations on how we set up the linear predictor follows the **Hedonic Price Model** (**HPM**)

Fit INLA

Data come from **Milan** Real Estate Rental within the municipality borders, extracted in **date: 22nd January 2021**

$$g(\mu_i) = \eta_i$$

$$\xi_i \sim N\left(\mathbf{0}, \mathbf{Q}_{\mathscr{C}}^{-1}\right)$$

$$\boldsymbol{\theta} = \{\boldsymbol{\xi}, \boldsymbol{\beta}\}$$

$$\mathbf{y} \sim \text{Normal}\left(\boldsymbol{\mu}, \sigma_e^2\right)$$

$$\psi_1 = \left\{\sigma_\varepsilon^2\right\}$$

The model Error

$$\psi_2 = \left(\sigma_\xi^2, \phi, \nu\right)$$

Variance GP

Hyper Params from Matèrn

Likelihood < *higher* >

- A **Normal distribution** function is specified on the response **Y,** i.e. the **log monthly price.** log transformation is required to normalize data.
- **sigma_e** is the measurement error precision hyper-param for **psi_1**
- The mean is equal to the linear predictor since link function **g()** is identity

Fit INLA

Data come from **Milan** Real Estate Rental within the municipality borders, extracted in **date: 22nd January 2021**

$$g(\mu_i) = \eta_i$$
$$\xi_i \sim N\left(\mathbf{0}, \mathbf{Q}_{\mathscr{C}}^{-1}\right)$$
$$\boldsymbol{\theta} = \{\boldsymbol{\xi}, \boldsymbol{\beta}\}$$

$$\psi_1 = \left\{\sigma_\varepsilon^2\right\}$$

The model Error

$$\psi_2 = \left(\sigma_\xi^2, \phi, \nu\right)$$

Variance GP

Hyper Params from Matèrn

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{\psi}_2) = \mathrm{MVN}(0, \boldsymbol{Q}(\boldsymbol{\psi}_2))$$

GMRF < *medium* >

- The **GMRF** is multivariate Normal centered in 0 with wit precision depending on parameters coming from the **regression coefficients** and the **Gaussian Process** treated with Matèrn becoming GMRF.
- This passage is critical
- The precision matrix **Q** depends on psi_2 which is the model error.

Fit INLA

Data come from **Milan** Real Estate Rental within the municipality borders, extracted in **date: 22nd January 2021**

$$g(\mu_i) = \eta_i$$
$$\xi_i \sim N\left(\mathbf{0}, \mathbf{Q}_{\mathscr{C}}^{-1}\right)$$
$$\boldsymbol{\theta} = \{\boldsymbol{\xi}, \boldsymbol{\beta}\}$$

$$\psi_1 = \left\{\sigma_\varepsilon^2\right\}$$

The model Error

$$\boldsymbol{\psi}_2 = \left(\sigma_\xi^2, \phi, \nu\right)$$

Variance GP

Hyper Params from Matèrn

$$\boldsymbol{\psi}_1 \sim N\left(0, 0.001^{-1}\right), \qquad l = 1$$
$$\psi_2 \sim N\left(0, 0.001^{-1}\right), j = \{1, 2, 3\}$$

Priors < *lower* >

Priors for hyper parameters are all set equal to **Gaussian Vagues** as the default choice for INLA since no expert knowledge is available.

Fit INLA

Data come from **Milan** Real Estate Rental within the municipality borders, extracted in **date: 22nd January 2021**

$$g(\mu_i) = \eta_i$$

$$\xi_i \sim N\left(\mathbf{0}, \mathbf{Q}_{\mathscr{C}}^{-1}\right)$$

$$\boldsymbol{\theta} = \{\boldsymbol{\xi}, \boldsymbol{\beta}\}$$

$$\psi_1 = \left\{\sigma_\varepsilon^2\right\}$$

The model Error

$$\boldsymbol{\psi_2} = \left(\sigma_\xi^2, \phi, \nu\right)$$

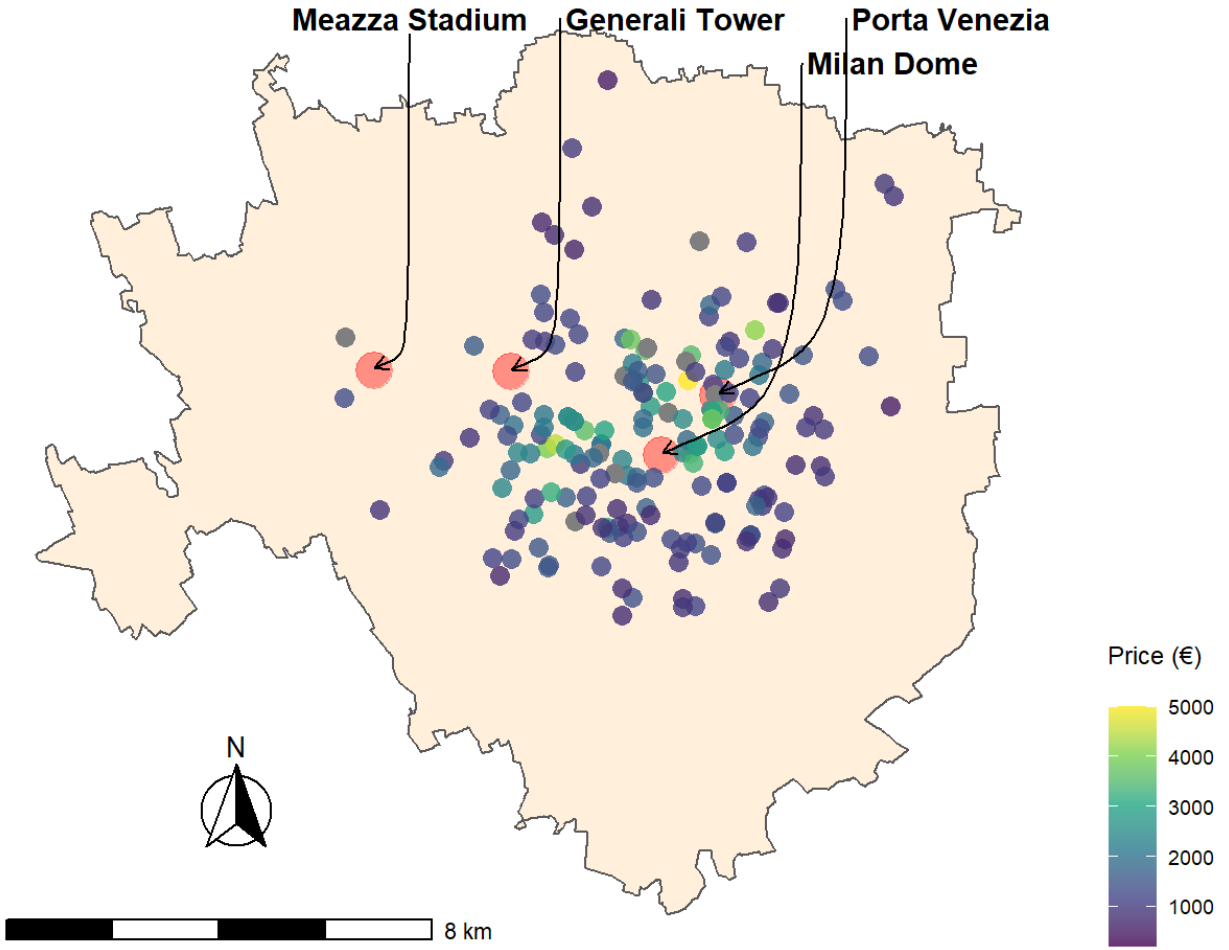
Variance GP

Hyper Params from Matèrn

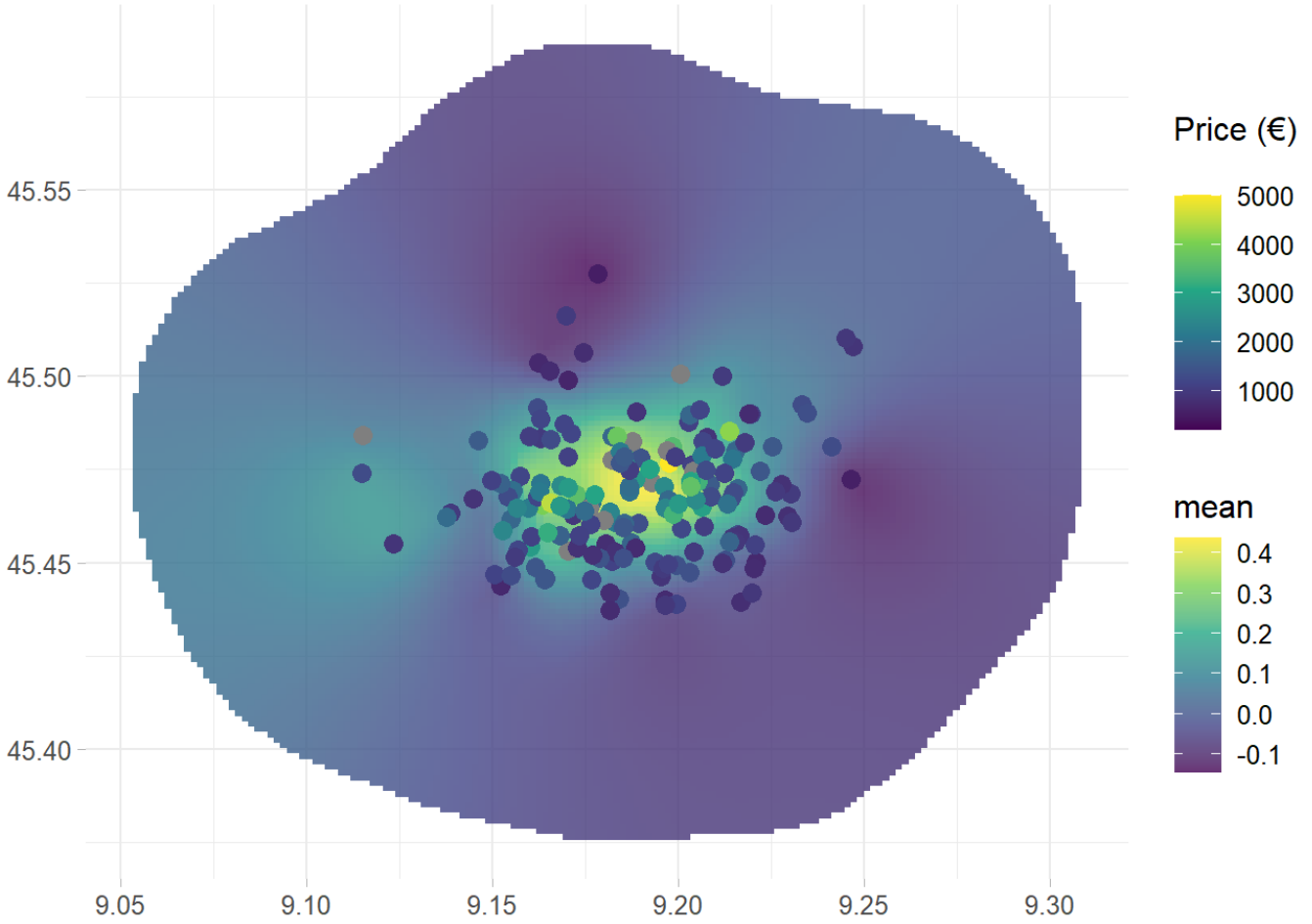
Laplace Approx.

Laplace is taken on teach element of the posterior marginal distribution of the latent field and the hyper parameter vector.

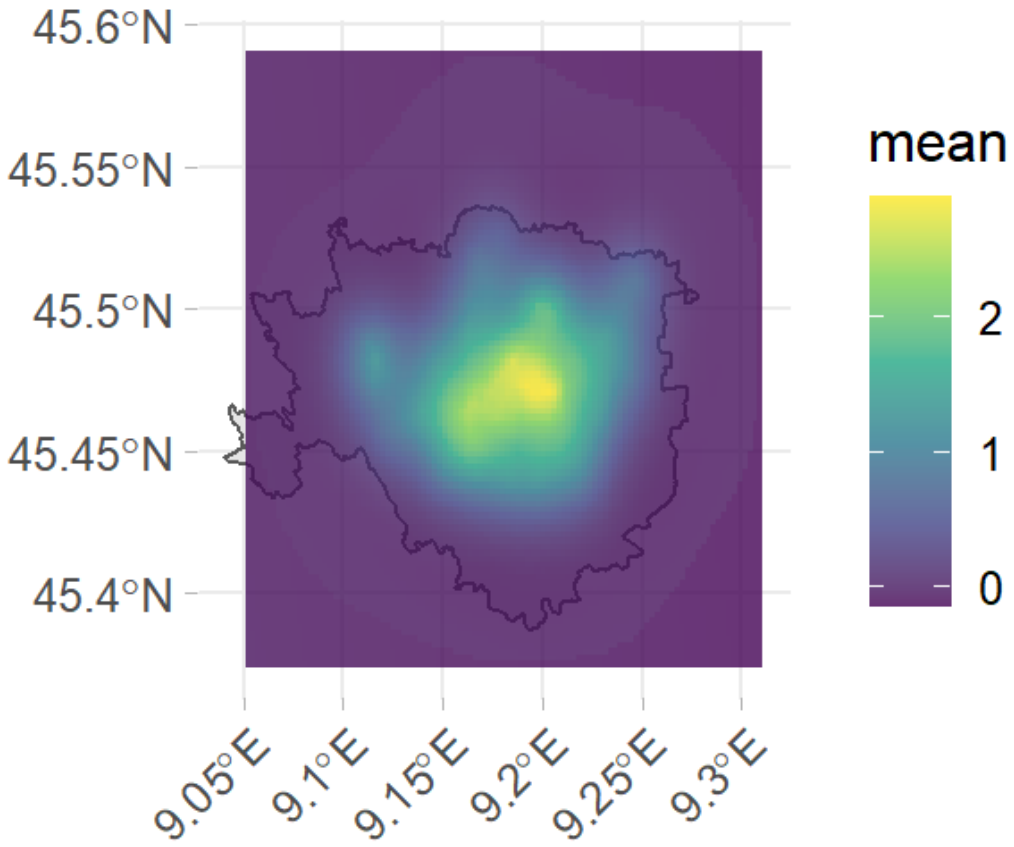
Data Glimpse

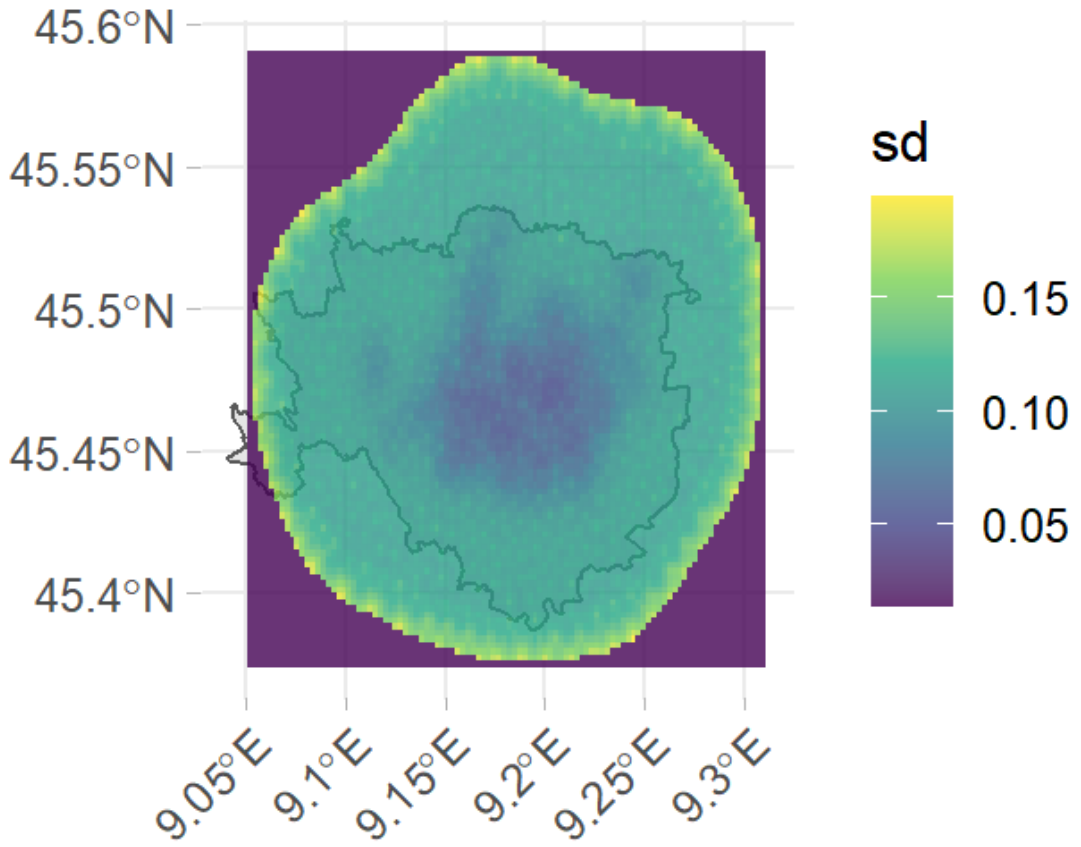


Plot the GMRF



Spatial Prediction





Main References

- **Spatial and Spatio-Temporal Bayesian Models with R-Inla**

  Blangiardo, Michela & Marta, Cameletti. 2015; Wiley.

- **Bayesian Inference with Inla. Chapman**

  Gómez Rubio, Virgilio. 2020; Hall/CRC. gitbook

- **Gaussian Markov Random Fields**

  Rue, Havard, & Leonhard Held. 2005; Chapman Hall/CRC.

- **Bayesian Regression Modeling with Inla**

  Wang Xiaofeng, Yu Ryan Yue, & Julian J Faraway. 2018; CRC Press

- **Spatio-Temporal Modeling of Particulate Matter Concentration Through the SPDE Approach**

  Cameletti, Michela, Finn Lindgren, Daniel Simpson, & Håvard Rue. 2012. AStA Advances in Statistical Analysis 97 (2): 109–31.

- **Advanced Spatial Modeling with Stochastic Partial Differential Equations Using R and Inla**

  Krainski, Elias T. 2019.; Chapman; Hall/CRC. gitbook

- **Geospatial Health Data**

  Moraga, Paula. 2019; Chapman; Hall/CRC. gitbook

QR code Ref

The thesis' website and API docs are completely hosted online and are browsable by scanning the following code.

Appendix I

Summary of the **model coefficients** arranged by descending mean, **top 12.**

Table 7.1: Summary statistics for the top 10 coefficients arranged by descending mean

coefficients	mean	sd	0.025quant	0.5quant	0.975quant
Intercept	5.46	12.91	-19.89	5.46	30.79
totlocaliTrilocale	1.19	12.91	-24.15	1.19	26.52
totlocali5+	1.18	12.91	-24.17	1.18	26.50
totlocaliBilocale	1.09	12.91	-24.25	1.09	26.42
totlocaliQuadrilocale	1.03	12.91	-24.32	1.03	26.36
totlocaliPentalocale	0.97	12.91	-24.38	0.97	26.29
heatingAutonomo, a pavimento, alimentato a pompa di calore	0.67	0.21	0.26	0.66	1.08
receptionyes	0.46	0.16	0.14	0.46	0.78
heatingCentralizzato, ad aria	0.33	0.17	0.00	0.33	0.66
heatingCentralizzato, a pavimento, alimentato a pompa di calore	0.31	0.15	0.01	0.31	0.60

Appendix II

Which are the **most profitable/valuable properties** in Milan w.r.t. a single square meter footage, depurated by outliers.

Table 6.2: The most profitable properties per single square meter footage at the date of2021-02-02

location	totlocali	price	sqfeet	floor	totpiani	abs_price
via della spiga 23	Bilocale	2500	55	2	4 piani	45.45€
via dei giardini C.A.	Multilocale	18500	415	2	6 piani	44.58€
piazza san babila C.A.	Bilocale	1833	42	1	4 piani	43.64€
ottimo stato nono piano, C.A.	Monolocale	2000	47	9	11 piani	42.55€
via tommaso salvini 1	Multilocale	15000	360	5	6 piani	41.67€
via cappuccini C.A.	Trilocale	4000	100	3	5 piani	40€