Manuel Acquistapace
Stefano Billeter

# Light pollution
# Analysis report

SYLLABUS

## 1. Introduction and Data Description

### Project Overview:

Our study delved into analyzing light pollution using Bayesian programming. Focusing on Mag/arcsec^2, a critical astronomical unit, the project aimed to quantify light pollution in various urban and natural settings. The unit represents the brightness of celestial objects per square arcsecond, with lower values indicating brighter sources.  We would like to thank Prof. Stefano Sposetti (Vice President of the Swiss Astronomical Society) and Ing. Stefano Klett (founder of Dark-Sky Switzerland, Italian language division) for their advice.

For further insights into convergence diagnosis and additional technical comments, please refer to the notebook.

### Data Characteristics:

The dataset encompassed measurements from various locations in Ticino. The data have been downloaded from https://www.oasi.ti.ch/web/esplora-dati/, as suggested. For suggestions regarding the priors' hyperparameters and the ROPE for the Hypothesis Testing, we derived our knowledge mainly from the following website: https://www.lightpollutionmap.info.

### Goal of the Report

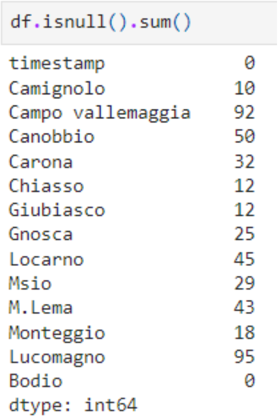
The primary objective of this report is to present a Bayesian analysis of luminescence data, focusing on understanding the effects of geographical and demographic factors on light pollution. By employing statistical models, including linear regression and hierarchical modeling, this report aims to uncover the patterns and relationships within the dataset, offering insights into the dynamics of luminescence across various locations in Ticino.

### Intended Audience

This report is designed for a range of audiences, including environmental scientists, data analysts, and policymakers. It is particularly valuable for professionals in environmental studies and urban planning who are interested in the impact of geographical features on light pollution. The report's findings are also relevant for academic researchers and students specializing in statistical modeling and environmental sciences, providing a practical application of Bayesian analysis in real-world data.

## 2. Descriptive Analysis

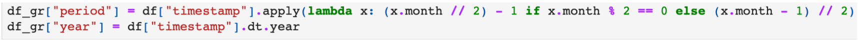
The descriptive analysis identified varying degrees of missing data. For instance, Campo Vallemaggia had 92 missing observations, and Lucomagno 95. Our decision to avoid imputing these missing values stemmed from the potential risk of introducing biases, which could skew the study's findings. By refraining from imputation, we maintained the dataset's integrity, ensuring that the analysis was based on observed data rather than estimates or assumptions.



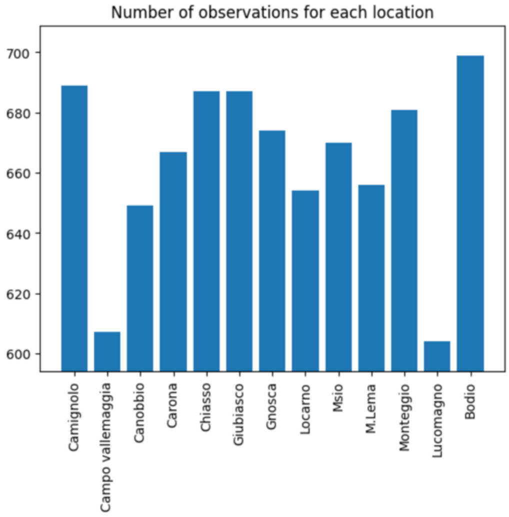
```
df.isnull().sum()

timestamp            0
Camignolo           10
Campo vallemaggia   92
Canobbio            50
Carona              32
Chiasso             12
Giubiasco           12
Gnosca              25
Locarno             45
Msio                29
M.Lema              43
Monteggio           18
Lucomagno           95
Bodio                0
dtype: int64
```

## 3. Data Grouping and Missing Value Management

The data was grouped into periods for a focused analysis. Due to different number of missing values, the daily observations count revealed variations among locations: Camignolo had 689 observations, while Lucomagno, the least observed location, had 604. Bodio, on the other hand, had the highest count with 699 daily observations.



```
df_gr["period"] = df["timestamp"].apply(lambda x: (x.month // 2) - 1 if x.month % 2 == 0 else (x.month - 1) // 2)
df_gr["year"] = df["timestamp"].dt.year
```



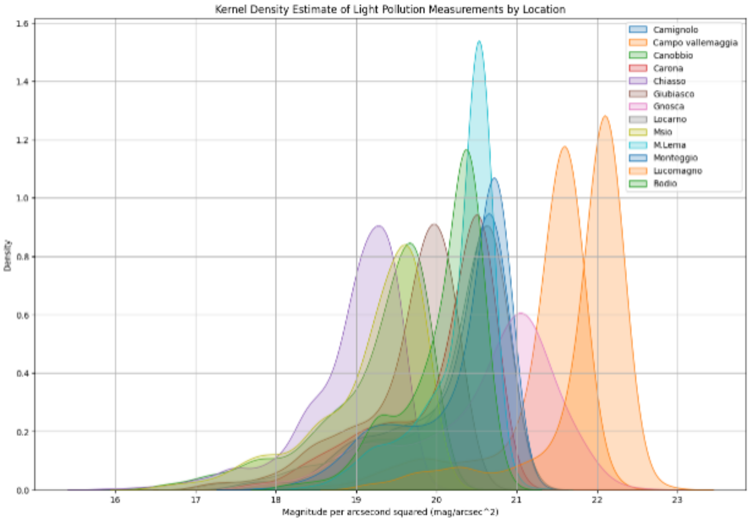
### Handling Missing Values:

Our approach to managing missing values involved converting the data into Python dictionaries. This method allowed for separate handling of NaN values, ensuring that only valid data points were considered in the analysis.

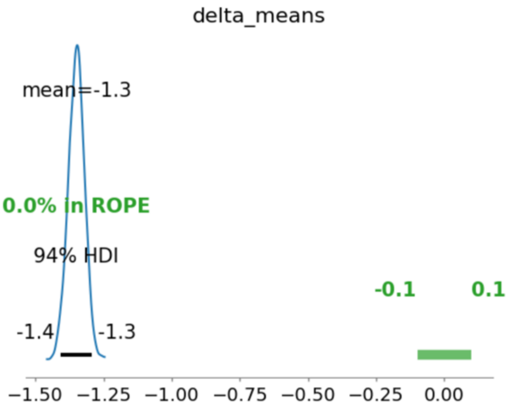
## 4. Kernel Density Estimates (KDE)

### Data Distribution Insights:

The KDE plots were a pivotal part of the analysis, revealing that while most data points followed a Gaussian distribution, significant outliers were present. This finding was crucial as it indicated areas with atypical light pollution levels, which could be due to environmental factors or specific urban developments. These outliers also justified the choice of T-Student distribution for the hypothesis testing, as it is more accommodating of such anomalies.



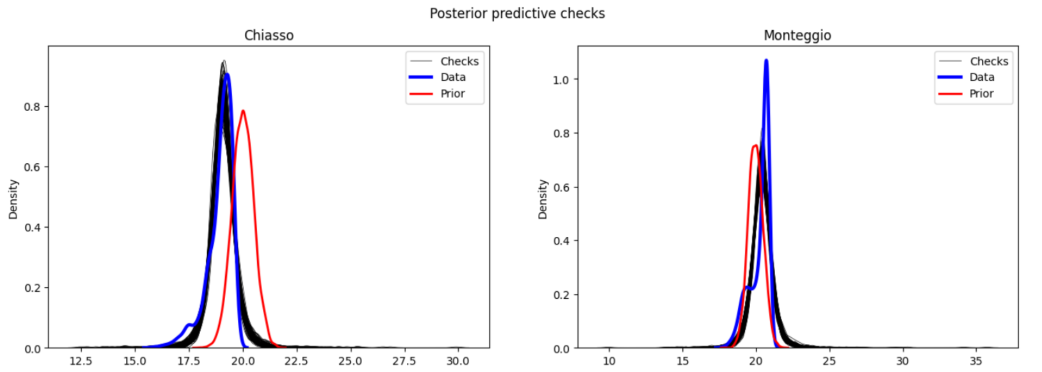
5. Hypothesis Testing (Chiasso vs. Monteggio)



In comparing Chiasso and Monteggio, the hypothesis testing revealed interesting insights. For instance, considering the mean light pollution for characterizing the distributions, in Chiasso was found to be 19.067, whereas in Monteggio, it was slightly higher at 20.417. The delta_means parameter, calculated as -1.350, indicated that Monteggio experienced higher mean light pollution than Chiasso. The ROPE is defined as the difference range between -0.1 and 0.1, given that, following the heatmap on the aforementioned website, the difference in luminescence between villages presenting similar features (inhabitants number, altitude…) falls withing this range.

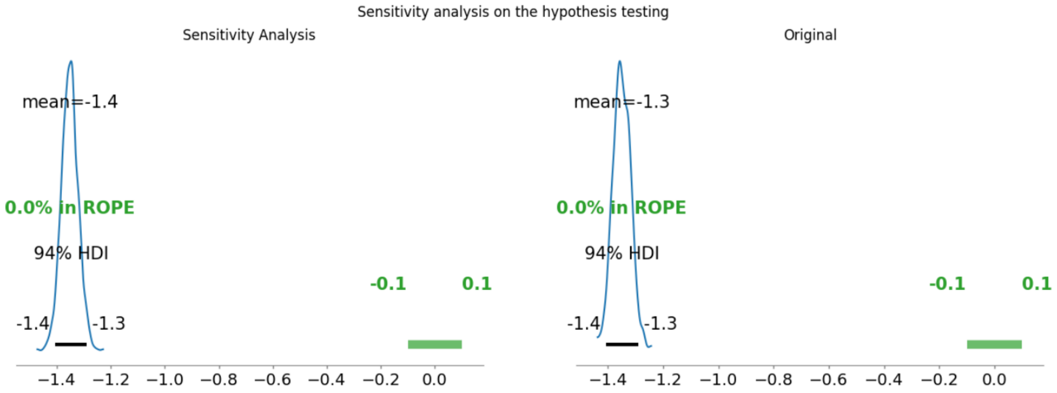
6. Posterior Predictive Checks (Chiasso and Monteggio)

The posterior predictive checks provided a robust validation of the Bayesian models. For Chiasso, the checks showed a high degree of alignment with the observed data, indicating a good model fit. In contrast, Monteggio displayed more outliers, which slightly shifted the model's predictions. Despite these discrepancies, the model fitting was reliable.

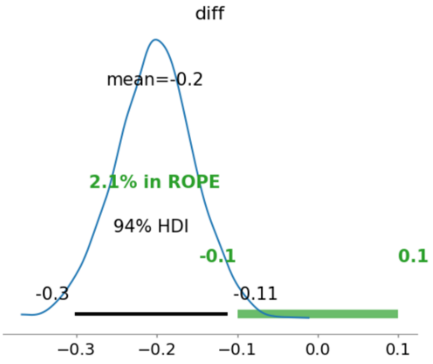


7. Prior Sensitivity Analysis

The sensitivity analysis played an important role in testing the robustness of the study's conclusions. By varying the prior distributions and observing the resulting changes in the posterior distributions, we could confirm the stability of the findings. This analysis revealed that the delta_means remained consistent across different prior assumptions, highlighting the reliability of the hypothesis testing results.

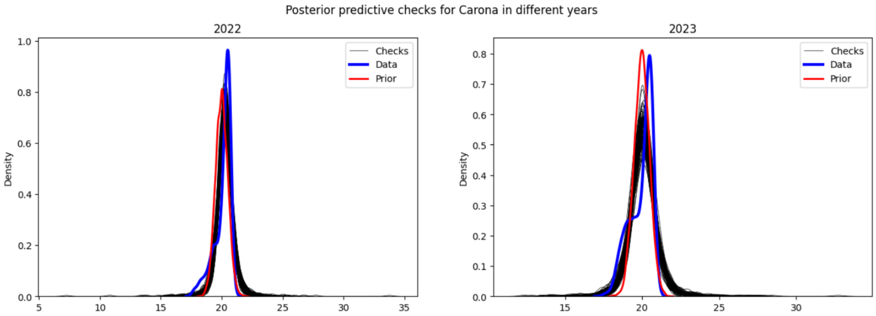


## 8. Hypothesis Testing and Posterior Predictive Checks (Carona 2022 vs. 2023)



The hypothesis test for Carona revealed a decrease in mean luminescence from 20.271 in 2022 to 20.068 in 2023. This change, while seemingly small, could indicate significant environmental or urban development impacting light pollution. The posterior predictive checks affirmed the model's accuracy in capturing these temporal changes. Although the larger area of overlap between the distribution of means difference and the ROPE, we can confidently “reject” the null hypothesis.

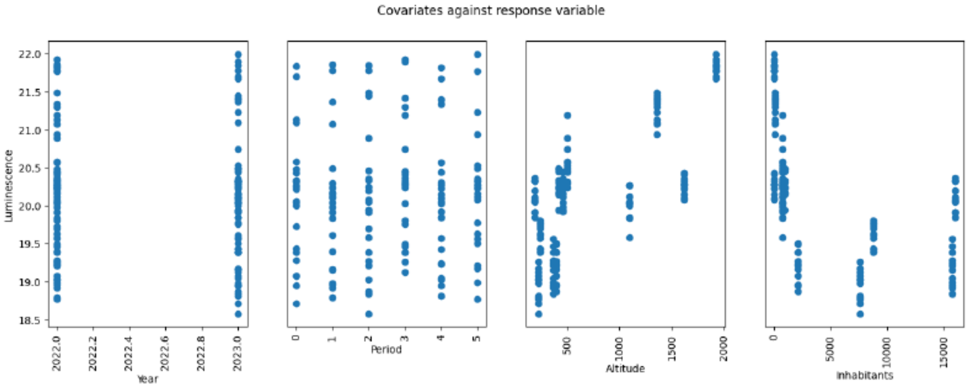
The posterior predictive checks below suggest that the model was able to understand the data. Moreover, we can clearly see the that our priors already well shaped the available data.



## 9. Linear Regression

### Data Preparation

In our Bayesian analysis, linear regression is employed to discern the relationship between luminescence and geographical factors. The dataset features luminescence readings from several locations with each location characterized by its altitude, year, and observation period. An initial scatter plot analyzes suggested correlations between luminescence and the considered covariates, highlighting the altitude as the only covariate.



The model is defined as follows:

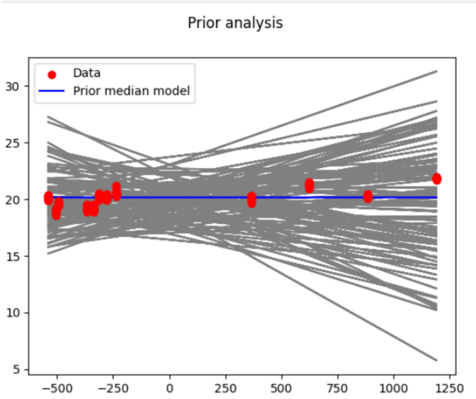
$$Model\ definition:$$

$$\beta_0 \sim N(\mu_{\beta_0}, \sigma_{\beta_0}),\ intercept$$

$$\beta_1 \sim N(\mu_{\beta_1}, \sigma_{\beta_1}),\ slope$$

$$\epsilon \sim HN(\xi)$$

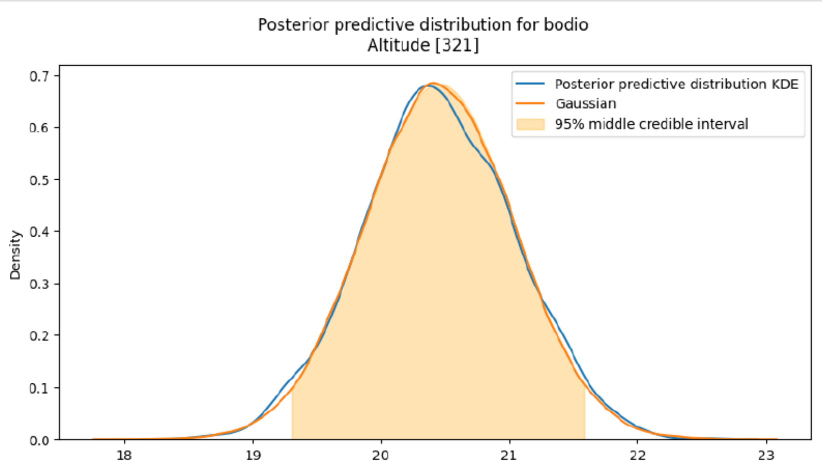
$$Y|\epsilon,\ \beta_0,\ \beta_1 \sim N(\beta_0 +\ \beta_1 * Altitude, \epsilon)$$



The data-driven priors are due to unavailable knowledge concerning the relationship between altitude and luminescence. The prior analysis plot on the right illustrates the influence of altitude on luminance, with the spread of gray lines representing possible regression lines drawn from the priors. Given the centering of the covariate, the prior median model goes through a point close to ¯y (20.11) for centered altitude values close to 0, as expected. Moreover, the large uncertainty of the priors derived from the gray lines reflects unavailable knowledge.

Results

The KDE plot shows the distribution of predicted luminance values, which closely follows the overlaid Gaussian curve, suggesting that the predictions are well-modeled by a normal distribution. The 95% middle credible interval highlighted in orange represents the range within which we can expect the true luminance values to fall with 95% probability. This interval, reflecting the model's uncertainty, is reasonably narrow, indicating a good level of precision in the predictions.



10. Hierarchical Model Analysis

In our Bayesian analysis, we delve into the Hierarchical Model to address the complexities inherent in our luminescence data. This approach is particularly adept at handling the multi-level nature of our dataset, which includes readings from various locations, each with its own unique altitude and demographic characteristics.

The Hierarchical Model in our study is designed to capture both the individual characteristics of each location and the overall trends across all sites. The model is structured with two levels:

Level 1 (Local Level): Models the luminescence readings at each location, considering local factors like altitude and population.

Level 2 (Global Level): Captures the broader trends and variations across different locations.

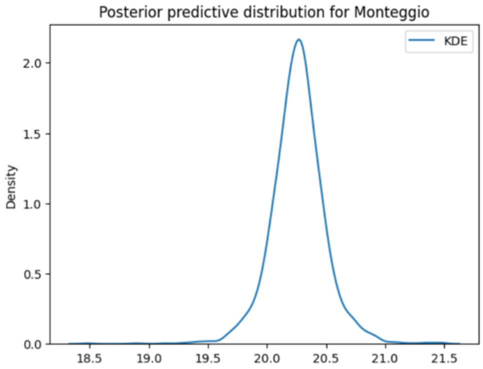
$$\mu_{brill} \sim N(20, 0.25)$$
$$\sigma_{brill} \sim HN(1)$$
$$\mu_i \sim N(\mu_{brill}, \sigma_{brill}) \;\; \forall i \in [0, 11]$$
$$\sigma_{loc} \sim HN(3)$$

*Equation 1 The data distribution is omitted here, but we modeled it with a TStudent with df 4 and sigma=sigma_loc and mu mu_loc_i*

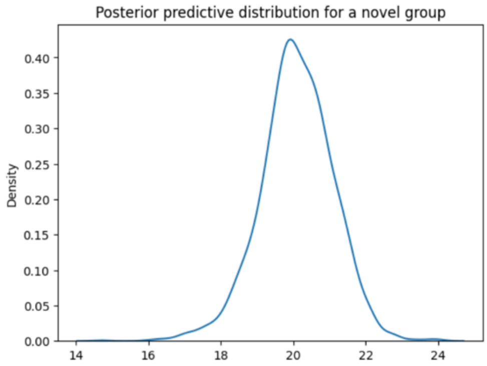
## 11. Posterior Predictive Distribution Analysis for Monteggio and Novel Groups

### Monteggio's Mean Luminance Estimation

In the last part of our analysis, we took into consideration the predictive distribution for Monteggio and a novel group. The mean luminance of Monteggio is estimated with a mean of approximately 20.26 and a standard deviation of 0.15. This precision in the mean estimation indicates a high level of confidence in the model's predictions for Monteggio.



The hierarchical model's robustness is further demonstrated through its application to a novel group, representing new data points or groups not explicitly covered in the training dataset.



The model extends to predict the luminescence for a novel group using a Kernel Density Estimation (KDE) plot. This approach utilizes the global luminance distribution (mu_brill, sigma_brill) and the distribution of the novel group to generate predictions.

The KDE plot of the posterior predictive distribution for a novel group exhibits higher uncertainty compared to the distribution for Monteggio. This increased uncertainty is attributed to the model considering both the global luminance distribution and the specific uncertainty related to the novel group.