ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/376270800

Exploring Seasonal Pollen Probabilities: A Comparative Approach

Preprint · December 2023

DOI: 10.13140/RG.2.2.33119.02728

CITATIONS
0

READS
65

3 authors, including:

Pareekshith Katti
Ambee
9 PUBLICATIONS **3** CITATIONS
SEE PROFILE

Nithin Srivatsav
Ambee
5 PUBLICATIONS **0** CITATIONS
SEE PROFILE

All content following this page was uploaded by Pareekshith Katti on 07 December 2023.

The user has requested enhancement of the downloaded file.

Exploring Seasonal Pollen Probabilities: A Comparative Approach

Pareekshith US Katti, N Nithin Srivatsav, Anmol Khilwani

Abstract

This research analyzed pollen count data from major urban centers such as Sydney, New York, Lyon, Paris, and London. Employing gamma and beta distributions, the study modeled the skewed nature of pollen counts and generated synthetic data to explore distribution characteristics over time. Focusing on weekly fluctuations for clearer seasonal patterns, the research evaluated accuracy using the R-squared score and extended findings to diverse cities. The study addressed challenges in creating synthetic datasets for pollen forecasting, emphasizing the importance of local adaptation, categorizing months based on pollen behavior, and scaling data using locally adapted averages. Demonstrated in Sydney, this scaling significantly improved alignment between synthetic and actual data. Beyond pollen analysis, the research underscored the relevance of considering local factors, seasonality, and data scaling in generating synthetic datasets for diverse regions. Contributing to the understanding of synthetic data generation, the study emphasized practical applications in environmental monitoring and public health. Ongoing research was essential for refining methods and enhancing accuracy across various domains.

Introduction

Pollen allergies affect a substantial 30% of the global population (*the-pollen-problem*, 2023), with managing chronic allergies costing an alarming $18 billion annually in the United States alone (*Facts and Stats - 50 Million Americans Have Allergies | ACAAI Patient*, 2022). Compounding this issue is a 96.3% surge in pollen levels since 1998, contributing to around 100 million yearly spikes in respiratory ailments in the US (*the-pollen-problem*, 2023). This upward trend in pollen concentration demands urgent attention due to its significant impact on both individual well-being and healthcare expenditure.

Traditional methods for measuring pollen levels involve manual collection and microscopic analysis. These include microscopic examination, where airborne particles are collected on surfaces and identified using light microscopes, Rotorod Sampler, a mechanical device with a rotating rod coated with a sticky substance for outdoor particle collection (Frenz et.al., 1997), Burkard Trap (Levetin, Estelle, et al. 2000), a widely used pollen sampler with a rotating drum coated in silicone grease, Pollen Slide Trap, which uses a sticky surface to collect particles transferred onto a glass slide for microscopic analysis and Gravimetric Methods, measuring the increase in weight on surfaces due to pollen deposition with a potential need for additional microscopic identification. Traditional methods have played a crucial role in understanding pollen distribution. However, these methods exhibit several shortcomings. Microscopic examination introduces subjectivity and the potential for human error in pollen identification, impacting the consistency and reliability of results. Additionally, traditional techniques are often limited in spatial coverage, providing localized information at specific monitoring stations and failing to capture the broader picture of pollen distribution across diverse landscapes.

Moreover, these methods can be time-consuming, labor-intensive, and may lack real-time data capabilities, hindering their scalability and effectiveness for large-scale or continuous monitoring. The dependency on weather conditions and the inability to offer quantitative precision further underscores the limitations of traditional pollen measurement approaches. A notable and concerning observation emerges as a significant deficiency in pollen monitoring infrastructure is identified in regions such as South America, Africa, and Asia.

To address these shortcomings, there is a need for a paradigm shift towards integrating modern technologies. Automated sampling devices, remote sensing tools, and predictive models (Papadogiannaki et.al., 2023) offer opportunities to enhance the accuracy, scalability, and timeliness of pollen-level assessments. These technologies mitigate the subjectivity associated with manual methods, provide broader spatial coverage, and enable real-time data collection. By combining the strengths of traditional and modern approaches, researchers can obtain a more comprehensive understanding of pollen dynamics, contributing to improved allergy management strategies and better-informed public health interventions.

The spatial coverage of traditional pollen monitoring methods is often limited due to inherent challenges such as stationary monitoring stations concentrated in urban areas, resource constraints hindering establishment in remote regions, logistical difficulties in expansive areas, and a dependency on population centers for monitoring (Buters et.al., 2018). This uneven distribution leads to gaps in coverage, leaving rural or less populated areas underrepresented. Additionally, traditional methods may not effectively capture dynamic changes in pollen distribution over time or consider the influence of diverse environmental conditions. To address these limitations and enhance spatial coverage, there is a growing recognition of the need to integrate modern technologies like remote sensing and predictive models (Picornell, A., et al.

2019), which offer a more comprehensive and dynamic approach to monitoring pollen levels across diverse landscapes.

Machine learning and probabilistic modeling present innovative solutions to address the lack of spatial coverage in pollen monitoring (Mills, Sophie A., et al. 2023). These technologies can predict pollen distribution by analyzing existing data and environmental variables and performing temporal and spatial interpolation by automating data processing and utilizing advanced algorithms, machine learning and probabilistic modeling offer scalable, cost-effective, and timely solutions to revolutionize pollen monitoring and bridge spatial gaps in regions where traditional infrastructure is limited.

Literature Review

A comprehensive examination delved into the pollen seasons across the North American region, encompassing both the United States and Canada. The study utilized data from 31 National Allergy Bureau (NAB) pollen stations spanning the years 2003 to 2017, with the primary objective of developing easily understandable pollen calendars catering to individuals with allergies and healthcare professionals. Noteworthy findings highlighted a negative correlation between the initiation date and duration of the primary pollen season. This underscored that locales experiencing earlier onset dates, particularly at lower latitudes, were associated with prolonged pollen seasons.

The research noted the widespread impact of pollen allergies on a significant segment of the U.S. population, emphasizing the crucial need for precise comprehension of the primary pollen season for accurate diagnosis and effective treatment. The study concluded by stressing the imperative for enhanced spatiotemporal monitoring of pollen concentrations while acknowledging the limitations in the coverage of NAB data. Recommendations included consistent year-round daily sampling and an expansion of monitoring stations. The study also emphasized the importance of documenting the spatial and temporal structure of the primary pollen season for allergenic pollen taxa.

Geographical nuances were explored, revealing a distinct latitudinal signal for the commencement date of pollen seasons, particularly concerning significant allergenic tree pollen taxa. Regional variations, such as the earlier onset of tree pollen seasons on the West Coast attributed to milder climates, were linked to factors like temperature and air transport from the Pacific Ocean. Notably, most locations experienced their tree pollen season from February to May, while the peak of grass pollen season occurred in June and July. Weed pollen, on the other hand, peaked in August and September (Lo, F., Bitz, C.M., Battisti, D.S. et al. 2019).

A study was conducted to construct a pollen calendar for Germany, exploring regional variations in pollen seasons across different parts of the country. The study specifically investigated the main flowering periods (MFP) of various tree and grass pollen types in the north, central, east, west, and south regions, utilizing data spanning from 2011 to 2016. Noteworthy findings included significant disparities in MFP, such as the grass MFP concluding 6 days earlier in the western region compared to both the northern and southern regions. The study underscored the importance of regional pollen calendars in providing accurate and region-specific information about pollen seasons, acknowledging the influence of climatic variations. According to the research, most trees had their main pollen flowering period between February and May, while grasses had it during June and July. Weed species had their main flowering period from late May to September (Werchan *et al.* 2019).

The University of Worcester in the United Kingdom published pollen calendars, which provided valuable insights into the seasonal dynamics and regional variations of allergenic pollen types. These calendars assisted individuals susceptible to pollen-induced allergies. The study detailed comprehensive pollen calendars for diverse regions, ranging from Scotland to Southeast England. It highlighted the appearance, risk levels, and peak periods of pollen types such as Hazel, Alder, Ash, Birch, Oak, Grass, and the Nettle family. This information was derived from a 10-year analysis (2003 to 2014) of main UK pollen allergens, offering a robust historical perspective on the prevalence and risk of allergenic pollen. Additionally, the inclusion of a generalized pollen calendar contributed to a holistic understanding of when major allergenic plants were in flower during the UK's pollen season, emphasizing the importance of considering geographical and temporal factors. The overall seasonality was similar to what had been observed in the pollen calendars published in Germany, with some

variations (*Pollen Calendars by Area - University of Worcester*, n.d.). Pollen calendar published in Ireland also showed similar seasonality (*Earlscliffe - Howth Weather*, n.d.).

A research presented a 15-year airborne pollen survey in Portugal, aimed at creating pollen calendars for seven monitoring regions. The study, conducted by the Portuguese Aerobiology Network, recorded 14 airborne pollen types, with 64.2% from trees, 28.5% from herbs, and 7.1% from weeds. Dominant allergenic types included Poaceae, Quercus spp., Urticaceae, and Cupressaceae. The average pollen index was higher in mainland Portugal than in the Islands, showing an increasing trend over the years, notably in Coimbra, Évora, and Porto. Grass pollen sensitization was prevalent among patients (34.4%), followed by Olea (21.3%) and Parietaria (17.5%). The study revealed a Mediterranean pollen spectrum, with varied prevalence in different regions. The pollen calendars indicated a concentration of allergenic taxa from March to July. The prevalence of respiratory allergies induced by pollen in Europe has been increasing. Despite seasonal variations, airborne allergens could occur throughout the year in specific regions, emphasizing the importance of continuous aerobiological research. The pollen calendar showed similar seasonality to Germany and the UK with some variations (Camacho, Irene, *et al.* 2020). Pollen calendars published in Spain also showed similar seasonality (*Estación Aerobiológica Universidad De Málaga*, n.d.).

A study conducted in urban areas of Australia and New Zealand revealed distinct shifts in pollen seasons, transitioning from prolonged periods in tropical regions to shorter durations in temperate zones, a phenomenon attributed to the influence of solar radiation incidence on pollen production. The research acknowledged the significant impact of surrounding land use on airborne pollen, noting variations in urban areas bordered by agricultural landscapes compared to those with greater forest cover. The study emphasized the potential for alterations in land use at urban boundaries to influence the types of airborne pollen within urban areas. Notably, cities such as

Sydney, Canberra, Melbourne, and Hobart exhibited overlapping pollen seasons for multiple species with some variations, highlighting the complexity of regional pollen dynamics in these urban environments (Haberle, Simon G., et al., 2014).

In a comprehensive analysis of airborne pollen in Chandigarh from 2018 to 2020, 74 pollen types were identified, and the Annual Pollen Integral was calculated. Morus alba emerged as the primary contributor with the highest annual mean pollen percentage. Two distinct pollen seasons were evident: spring (February–April), dominated by arboreal pollen types, and autumn (August–October), characterized by herbaceous pollen types. The pollen calendar illustrated notable concentrations and extended season lengths over the two years (Ravindra, Khaiwal, et al. 2021). A parallel study in Allahabad revealed two main pollen seasons: February–May, featuring arboreal taxa as chief contributors, and September–October, dominated by grasses. Correlation analyses highlighted negative relationships between daily pollen counts and minimum temperature, relative humidity, and rainfall (Sahney et al., 2008). Both investigations underscored the significance of comprehending local pollen dynamics for managing pollen-related allergic diseases, emphasizing consistent seasonal patterns between the two cities.

Research on airborne pollen concentrations in Argentina, specifically in Bariloche, Cordoba, Bahia Blanca, and Santa Rosa, revealed a distinct seasonality, with peak pollen concentrations observed in spring, particularly in August and October. Tree pollen dominated across all locations, followed by grass and weed pollen. The study underscored the impact of airborne pollen on seasonal allergic diseases in Argentina, highlighting a significant scarcity of literature on the prevalence of allergic diseases and airborne allergens in the country. Despite overlapping pollen seasons in most cities during September-October, Bahia Blanca experienced its peak in August. The research called for further investigation into pollen seasonality, its link to allergic diseases,

patient sensitization, climate change effects, standardization of pollen concentrations, and the development of a national pollen map (Ramon et al., 2020).

The significant challenge lies in the scarcity of prevailing literature and data. However, the presence of overlapping seasonalities offers an opportunity to employ probability models for the analysis of pollen data.

Methodology

Data for this research were acquired for three years from diverse urban locations. The primary dataset encompassed time-stamped pollen count records for each city. To gain preliminary insights into the distribution patterns of pollen counts across various cities and temporal intervals, we employed descriptive statistical measures and data visualization techniques. This was repeated five times, each instance utilizing distinct seed values.

The primary goal of our analysis was to construct models for the probability distributions of pollen counts in various cities. To accomplish this, we utilized Beta and Gamma distributions, recognized for their suitability in capturing the fluctuations in pollen counts. Subsequently, the parameters of these distributions were employed to simulate pollen counts. The overall count was computed by aggregating individual pollen counts.

The study encompassed five distinct urban locales: Sydney, New York, Lyon, Paris, and London. For each city under investigation, histograms with kernel density estimations (KDEs) were produced independently. The findings of the analysis revealed a positively skewed distribution, primarily concentrated towards lower pollen counts across all the urban centers studied.

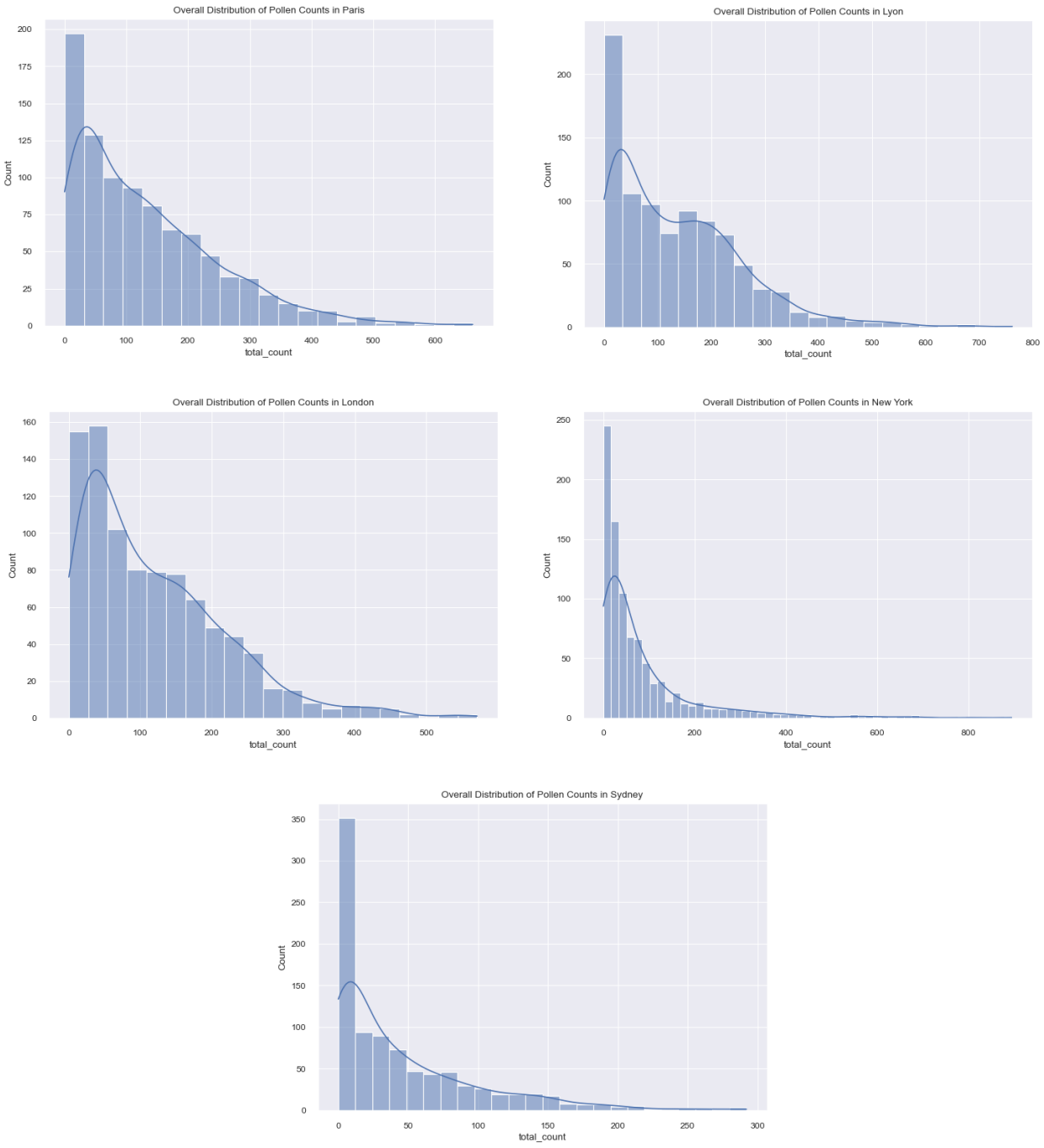

Fig 1: Distribution of Pollen Counts for the Selected Cities

The data underwent fitting procedures for both Gamma and Beta distributions. The rationale behind exploring these distributions stemmed from the inherent right-skewed characteristics exhibited by the data. The primary objective was to adeptly capture and

model this skewed data distribution. The research endeavor aimed to ascertain the most suitable distribution that concurred with the observed right-skewed pattern evident in the pollen count data.

Two distinct functions were developed to periodically fit gamma and beta distributions to the data, contingent upon the designated time intervals, such as months. With each unique time interval value, the corresponding pollen count values were isolated, and a distribution was fitted accordingly. The resultant distribution parameters were gathered and recorded for each time period.

Our methodology comprised several phases. To begin, we calculated distribution parameters customized to the Paris dataset. Leveraging these estimated distribution parameters, we synthesized data points to emulate the distribution traits observed during particular months or weeks, with a primary emphasis on preserving data integrity.

To assess the effectiveness of gamma and beta distributions in capturing the weekly fluctuations in pollen counts, and where applicable, the monthly variations, we quantified the accuracy of these approximations using the R-squared (R2) score. This score was calculated for both the actual data and the synthetic data, specifically for weekly aggregations. This evaluation provided insights into the degree of concordance between the distributions and the observed patterns of pollen counts in Paris over both monthly and weekly timeframes.

Our decision to concentrate on weekly fluctuations instead of daily ones was motivated by the substantial variance apparent in daily data. Analyzing daily variations would introduce significant noise and variability, rendering it difficult to discern underlying patterns or seasonality in the pollen count data. By aggregating the data weekly, we intended to minimize the influence of daily noise, thereby facilitating a more lucid

exploration of stable and consistent trends in pollen counts over time. This approach enabled us to effectively investigate seasonal patterns and variations while mitigating the complications associated with the daily data's fluctuating nature.

We evaluated the beta distribution's capability to replicate weekly fluctuations in pollen counts within the city of Paris. Following this assessment, we expanded our analysis to create synthetic data using the beta distribution parameters obtained from Paris. We then compared this synthetic data with the actual data from other cities situated within the same country, the same region, and even across a different hemisphere. Our goal was to evaluate the degree to which the beta distribution parameters derived from Paris effectively mirrored the genuine weekly patterns of pollen counts in these other cities.

For our next experiment, we categorized months into different levels based on the mean monthly total pollen count. This categorization aimed to align the seasonal patterns of Paris with those of New York and Sydney, facilitating a comparative assessment of pollen count dynamics. We estimated parameters describing the distribution characteristics at regular intervals. Subsequently, we generated synthetic data that adhered to these distribution patterns, making necessary adjustments to eliminate negative values. The data was then aggregated weekly for temporal analysis, and we quantified the alignment between observed and synthetic pollen counts using the R-squared score.

To ensure representative sampling, we implemented a stratified sampling method. This involved selecting 25% of the data from the city of Paris systematically to compute monthly means. We then scaled each month's dataset by its corresponding monthly mean. This process included grouping the data by month, selecting a random 25% sample within each month, and calculating monthly means by grouping the data again by month. Scaling factors were determined by dividing the monthly mean of observed

pollen counts by the monthly mean of synthetic pollen counts. The synthetic data underwent adjustments by scaling them with the scaling factors obtained earlier. Subsequently, we aggregated the data weekly to calculate the weekly averages for both observed and adjusted synthetic pollen counts. Finally, we computed an R-squared (R2) score to quantify the degree of alignment between the weekly averages of observed pollen counts and the adjusted synthetic pollen counts.

Our research methodology maintained consistency across New York City (NYC) and Sydney, Australia, following the same analytical procedures applied in Paris. Both cities' datasets underwent categorization based on predefined criteria. Synthetic data, generated to adhere to the beta distribution, were created for each city, utilizing beta distribution parameters initially estimated using Paris's data. These synthetic datasets were then adjusted to ensure non-negativity. The analysis centered on calculating weekly averages and assessing alignment with observed pollen counts through the R-squared (R2) score. Additionally, scaling factors were computed for each city based on monthly means, allowing for the normalization of synthetic data to reflect observed data dynamics.

Results

The study investigated the appropriateness of utilizing two distinct statistical distributions, namely the gamma and beta distributions, for representing the seasonal variation in pollen data. To assess the goodness of fit for these models, the research employed the R2 score, which offers an overall assessment of the extent to which the distributions aligned with the actual data.

Trial	Gamma (Monthly)	Beta (Monthly)	Gamma (Weekly)	Beta (Weekly)
Seed 0	0.42	0.70	0.71	0.86
Seed 1	0.39	0.69	0.70	0.74
Seed 2	0.39	0.66	0.72	0.78
Seed 3	0.42	0.69	0.70	0.77
Seed 4	0.44	0.70	0.69	0.76
Average	**0.412**	**0.688**	**0.704**	**0.782**

Table 1: Weekly R2 Scores for various trials for different probability distribution configurations

When assessing how well the gamma distribution fit the monthly pollen data, the R2 scores fell within the range of 0.39 to 0.44. These scores suggest a moderate level of fit. Nonetheless, it is worth highlighting that the gamma distribution encountered difficulties when it came to adequately modeling specific months, as evidenced by the lower R2 scores, especially those hovering around 0.39.

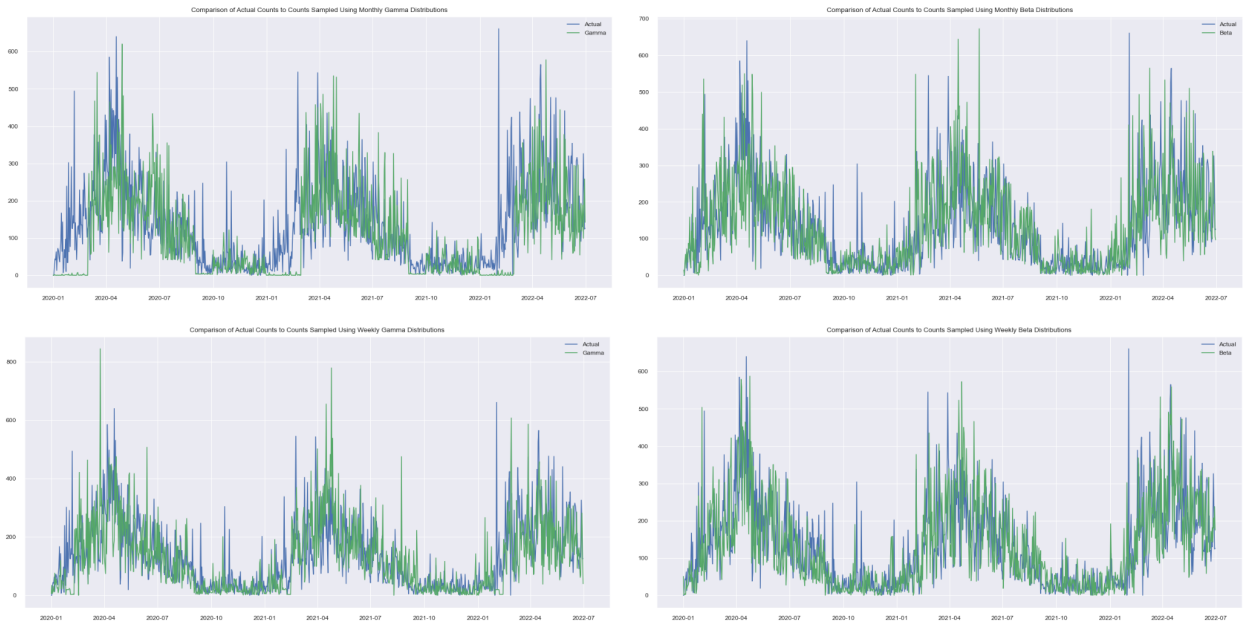

Fig 2: Comparison of Gamma and Beta Distributions under Monthly and Weekly Configurations (Seed 0)

Furthermore, when examining visual representations of the synthetic data from monthly gamma distributions, anomalies, and irregularities in trendlines were evident for certain months, pointing to limitations in its ability to accurately capture the subtleties present in the pollen data.

In stark contrast, the beta distribution consistently outperformed the gamma distribution in both monthly and weekly analyses. When we delved into the monthly distributions, the R2 scores for the beta distribution consistently fell within the range of 0.66 to 0.70, indicating a substantially stronger fit to the data. This pattern persisted when scrutinizing the data at a weekly granularity, with R2 scores ranging between 0.74 and 0.86. These results emphasize the beta distribution's reliability in representing seasonal pollen probability distributions, positioning it as a superior choice when compared to the gamma distribution. Its ability to yield consistently higher overall R2 scores underscores its capacity to offer a more precise depiction of the seasonal fluctuations in pollen data, thereby affirming its suitability for this research context.

Consequently, the beta distribution emerges as a more robust tool for comprehending and modeling pollen distribution patterns throughout the year.

The core focus of our analysis revolved around the utilization of the beta distribution parameters that we derived from the pollen data in Paris. We created a distribution for each week. We used these parameters to generate synthetic data, effectively simulating pollen distributions. To assess the reliability and applicability of this synthetic data in a broader context, we compared it to actual pollen data collected from several distinct cities, including Lyon, London, New York, and Sydney. This comparison provided us with valuable insights into how well our synthetic data performed when applied to various geographical locations. It allowed us to gauge the effectiveness of our modeling approach and draw meaningful conclusions about the generalizability of the beta distribution parameters obtained from Paris's data to other urban environments.

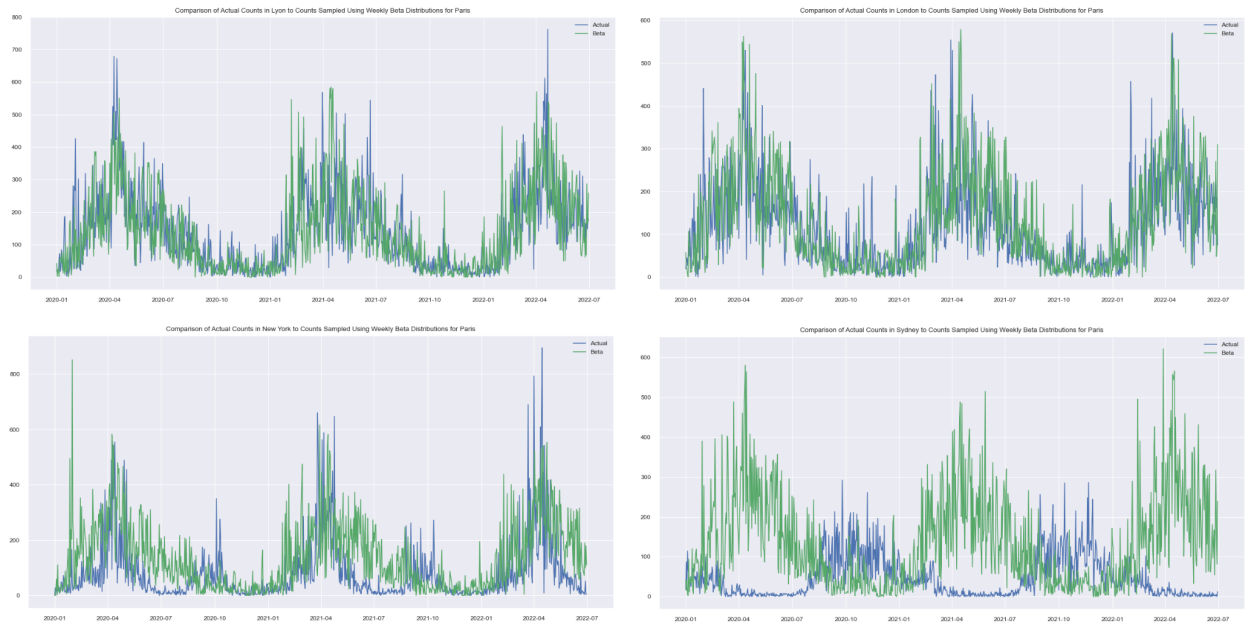

Fig 3: Comparison of Actual Counts for Other Cities vs Synthetic Data Generated Using Weekly Beta Distributions for Paris (Seed 0)

Trial	Lyon	London	New York	Sydney
Seed 0	0.72	0.61	-0.10	-12
Seed 1	0.73	0.70	-0.08	-12
Seed 2	0.74	0.56	-0.12	-12
Seed 3	0.77	0.58	-0.14	-11.60
Seed 4	0.72	0.48	-0.08	-12

Table 2: Weekly R2 Scores for different cities when compared against data generated from weekly distribution for Paris

As anticipated, the synthetic data yielded the least favorable results when applied to Sydney, a city situated in the Southern Hemisphere, which experiences opposite seasonal variations compared to Paris in the Northern Hemisphere. In all five trials conducted, Sydney consistently displayed notably low R2 scores, hovering around -12. This outcome aligns perfectly with our expectations, as it is well-established that the seasonal patterns in pollen data can diverge significantly between hemispheres due to the contrasting nature of their seasonal variations.

New York, another city located in the Northern Hemisphere like Paris, exhibited relatively low R2 scores, albeit with some variability observed across the different trials. These findings imply that, although there might be certain resemblances in seasonal patterns between Paris and New York, the synthetic data failed to capture and replicate these patterns accurately. The variation in R2 scores across trials underscores the nuanced and complex nature of pollen distribution, which can be influenced by local factors and conditions specific to each city, even within the same hemisphere.

In a similar geographic region to Paris, London exhibited a broad spectrum of R2 scores, ranging from 0.48 to 0.70 across the trials. This considerable variability in scores suggests that the synthetic data had some success in partially representing the seasonality found in London's pollen data. However, the diversity in these scores also indicates that the synthetic data may not have been able to precisely replicate the exact nuances of London's pollen distribution in every trial. It highlights the complexity of modeling pollen patterns even in closely related regions, underlining the potential impact of local factors and unique environmental conditions on pollen data variations.

Lyon, situated within the same country as Paris, showcased the highest R2 scores among the cities, spanning from 0.72 to 0.77. These results indicate that the synthetic data demonstrated relatively strong performance in replicating the seasonal patterns observed in both Paris and Lyon. This outcome aligns with expectations, given their close geographical proximity and shared regional characteristics. It underscores the utility of synthetic data in capturing and reproducing pollen distribution patterns in areas with similar environmental conditions and seasonal trends.

The performance of the synthetic data, generated using the beta distribution parameters, exhibited variations when applied to different cities. It demonstrated strong performance for cities within the same country or region, while its effectiveness diminished for cities located in different hemispheres. This observation underscores the critical importance of accounting for local environmental factors and geographical distinctions when undertaking pollen distribution modeling. It reinforces the idea that successful modeling should be tailored to the specific characteristics and seasonal variations unique to each geographic location.

To harmonize the seasonal patterns of pollen data across diverse regions, an alternative approach was adopted. Monthly data from Paris, New York, and Sydney were stratified according to percentile values corresponding to their respective monthly averages. The

rationale behind this strategy was to cluster months with analogous pollen patterns into the same category, acknowledging that although pollen seasonality may fluctuate monthly, the broader pollen behavior might remain stable during peak and off-seasons. This classification process yielded unique groupings for each city, segregating months into categories denoting low, moderate, high, or very high pollen behavior.

Month	Paris	New York	Sydney
1	Low	Low	Moderate
2	Moderate	Moderate	Moderate
3	High	High	Low
4	Very High	Very High	Low
5	High	High	Low
6	Moderate	Low	Low
7	Moderate	Low	Low
8	Low	Low	Moderate
9	Low	Low	High
10	Low	Moderate	High
11	Low	Low	Very High
12	Low	Low	Moderate

Table 3: Categorization of Months into Low, Moderate, High, and Medium based on pollen counts

After the categorization process, an analysis was undertaken using the weekly R2 score, a metric designed to gauge the degree of alignment between the synthetic and actual pollen data. While this categorization did address some of the seasonality concerns, a new challenge emerged - a noticeable variation in the magnitude of pollen counts across the different cities. To tackle this scaling discrepancy, a stratified sampling approach

was implemented. For each month, a random 25% sample of the data was utilized to compute monthly averages for each city. These computed averages were then employed to adjust the scale of the synthetic data for each corresponding month, effectively harmonizing it with the distinct ranges observed in each city's actual pollen data.

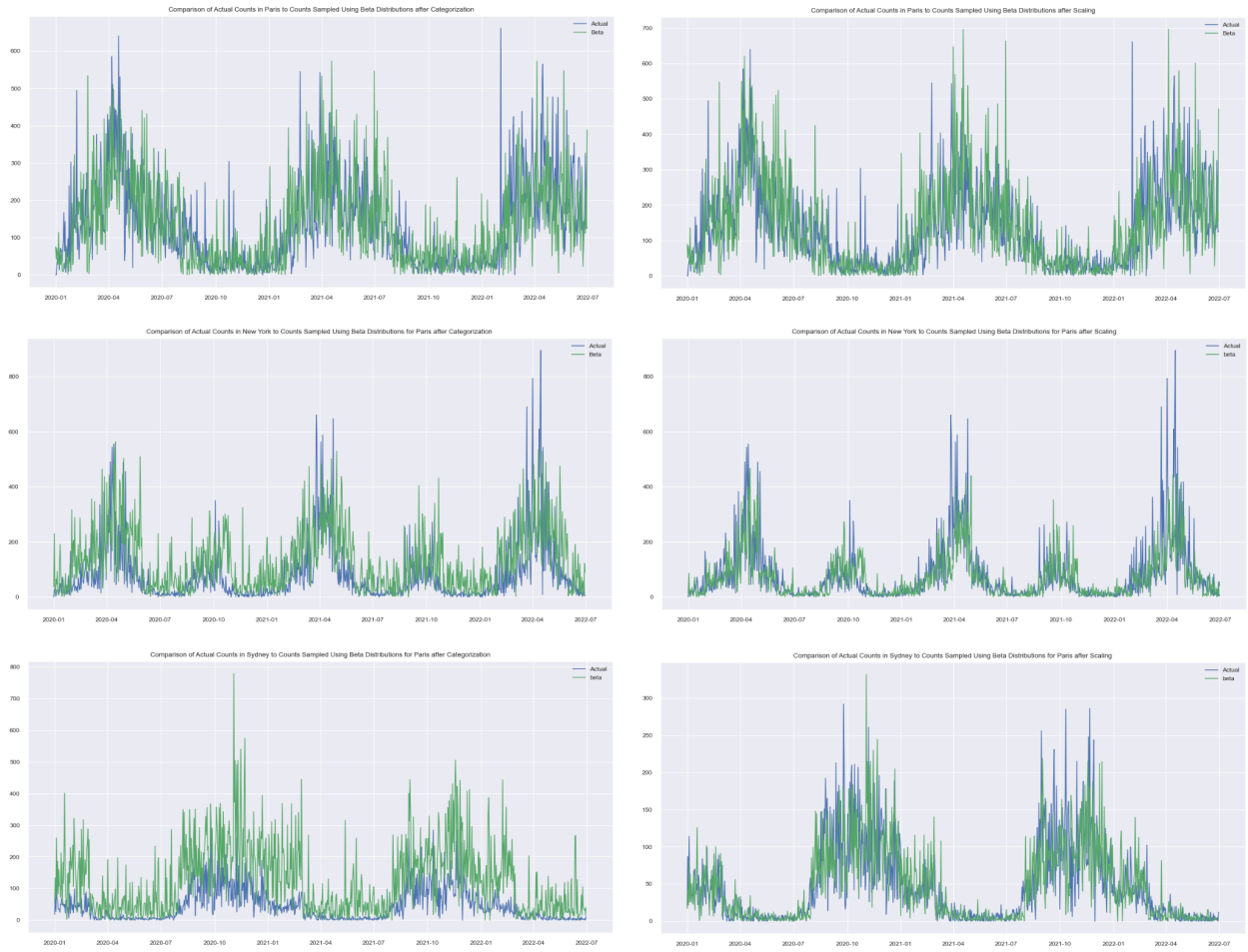

Fig 4: Comparison of Actual Counts and Synthetic Data After Categorization and Scaling (Seed 0)

Trails	Paris Before Scaling	Paris After Scaling	New York Before Scaling	New York After Scaling	Sydney Before Scaling	Sydney After Scaling
Seed 0	0.68	0.69	0.28	0.68	-3.7	0.79
Seed 1	0.64	0.60	0.25	0.71	-3.41	0.802
Seed 2	0.68	0.64	0.16	0.65	-2.98	0.80
Seed 3	0.71	0.75	0.20	0.73	-6.1	0.81
Seed 4	0.63	0.49	0.33	0.67	-3	0.76

Table 4: Weekly R2 Scores after categorization with and without scaling

Before scaling, Paris's data demonstrated a relatively strong fit between the synthetic pollen data, generated based on its distribution, and the actual pollen counts, with R2 scores ranging from 0.68 to 0.71 across five trials (Seeds 0 to 4). This indicated a close resemblance between the initial synthetic data and the actual pollen counts in Paris. Following the implementation of the scaling process, there were alterations in the R2 scores, with values spanning from 0.49 to 0.75. While certain trials exhibited improvements, as exemplified by Seed 3, where the R2 score increased to 0.75, others displayed variability, as observed in Seed 4 with a score of 0.49. In summary, scaling seemed to enhance the alignment of synthetic data with Paris's actual pollen data in specific instances, although the impact varied across different trials.

New York's data exhibited varying levels of fit before scaling between the synthetic pollen data, generated based on Paris's distribution, and the actual pollen counts. The R2 scores spanned from 0.16 to 0.33 across five trials (Seeds 0 to 4), indicating a lack of close alignment between the initial synthetic data and New York's actual pollen counts.

However, after implementing the scaling process, significant enhancements were evident. The R2 scores post-scaling ranged from 0.65 to 0.73, with the higher values signifying a substantially improved fit between the scaled synthetic data and the actual pollen counts. This underscored the effectiveness of the scaling process in bringing the synthetic data into better alignment with New York's pollen data, highlighting the pivotal role played by locally adapted monthly averages in achieving this improved fit.
Sydney's data initially exhibited negative R2 scores, ranging from -3.7 to -2.98 across five different trials (Seeds 0 to 4) before the scaling process was applied. These negative R2 scores pointed to a significant lack of alignment between the synthetic pollen data, generated based on Paris's distribution, and the actual pollen counts in Sydney. The synthetic values were notably higher than the actual values, indicating a substantial disparity. However, after implementing the scaling process, a remarkable improvement was observed. The R2 scores post-scaling ranged from 0.76 to 0.81, showcasing a strong fit between the scaled synthetic data and the actual pollen counts. This scaling procedure effectively addressed the issue of overestimation, bringing the synthetic data into much closer agreement with Sydney's actual pollen data and resulting in highly favorable R2 scores. This indicates the successful calibration of the synthetic data to the local context in Sydney.

The substantial increase in R2 scores following the scaling of the synthetic data in Sydney can be attributed to the successful adaptation of the synthetic dataset to the unique pollen characteristics of the city. Initially, the synthetic data generated using Paris's distribution resulted in negative R2 scores, indicating a significant overestimation of pollen counts in Sydney. However, the scaling process effectively resolved this issue by aligning the synthetic data with locally adjusted monthly averages, thereby matching the scale and seasonality of pollen counts in Sydney more accurately. This scaling procedure effectively mitigated the problem of overestimation,

reducing the disparity between the synthetic and actual data. This notable improvement underscores the critical importance of considering local factors and calibration when generating synthetic data for different regions, ultimately resulting in a much closer fit between the synthetic and actual pollen counts in Sydney.

This highlights that while addressing seasonality was an important advancement, ensuring the precise scaling of synthetic data using locally adapted monthly averages was vital to obtain accurate results in diverse geographic regions.

This approach highlights the significance of both capturing seasonality and adeptly scaling synthetic data when dealing with pollen data from various cities. It enables a more precise depiction of pollen behavior across regions, ultimately enhancing the alignment between synthetic and actual data, and yielding valuable insights into pollen forecasting and analysis.

Conclusion

In this research, we analyzed the generation of synthetic pollen data and its correlation with real pollen counts across diverse geographical regions. The experiment unveiled critical insights into the challenges and solutions related to synthesizing data for time-series analysis, specifically within the domains of pollen forecasting and environmental health.

Our investigation underscored the paramount importance of local adaptation in crafting synthetic datasets. While applying probability distributions yielded positive outcomes for similar regions, it became evident that simply modeling after one location's data was insufficient to capture the nuances of pollen behavior across diverse regions. Instead, our approach focused on categorizing months based on pollen behavior, maintaining seasonality, and addressing the significant challenge of scale mismatch.

The findings demonstrated that scaling synthetic data using locally adapted monthly averages played a crucial role in mitigating disparities between synthetic and actual data. Notably, a remarkable improvement was observed in the alignment of synthetic data with actual pollen counts in Sydney, where the initial synthetic data showed negative R2 scores. The scaling process rectified overestimation issues, bringing the synthetic data into close alignment with the local context, as evidenced by highly favorable R2 scores.

References

1. B. (2023, August 15). *the-pollen-problem*. https://www.bayer.com/en/news-stories/the-pollen-problem
2. *Facts and Stats - 50 Million Americans Have Allergies | ACAAI Patient*. (2022, September 16). ACAAI Public Website. https://acaai.org/allergies/allergies-101/facts-stats/
3. Lo, F., Bitz, C.M., Battisti, D.S. *et al.* Pollen calendars and maps of allergenic pollen in North America. *Aerobiologia* **35**, 613–633 (2019). https://doi.org/10.1007/s10453-019-09601-2
4. Werchan, M., Werchan, B. & Bergmann, KC. German pollen calendar 4.0: update of the regional pollen calendars 4.0 with measurement data for the period 2011–2016. *Allergo J Int* **28**, 160–162 (2019). https://doi.org/10.1007/s40629-019-0095-1
5. *Pollen Calendars by area - University Of Worcester*. (n.d.). https://www.worcester.ac.uk/about/academic-schools/school-of-science-and-the-environment/science-and-the-environment-research/national-pollen-and-aerobiology-research-unit/pollen-calendar.aspx
6. *Earlscliffe - Howth Weather*. (n.d.). https://weather.earlscliffe.com/wxpollen.php. Retrieved November 27, 2023, from https://weather.earlscliffe.com/wxpollen.php
7. Camacho, Irene, et al. "Airborne Pollen Calendar of Portugal: A 15-Year Survey (2002–2017)." Allergologia et Immunopathologia, vol. 48, no. 2, Mar. 2020, pp. 194–201. DOI.org (Crossref), https://doi.org/10.1016/j.aller.2019.06.012
8. *Estación aerobiológica Universidad de Málaga*. (n.d.). http://www.aerobiologia.uma.es/estaciones/teatinos.html
9. Haberle, Simon G., et al. "The Macroecology of Airborne Pollen in Australian and New Zealand Urban Areas." PLOS ONE, vol. 9, no. 5, May 2014, p. e97925. PLoS Journals, https://doi.org/10.1371/journal.pone.0097925
10. Ravindra, Khaiwal, et al. "Pollen Calendar to Depict Seasonal Periodicities of Airborne Pollen Species in a City Situated in Indo-Gangetic Plain, India." *Atmospheric Environment*, vol. 262, Oct. 2021, p. 118649. *ScienceDirect*, https://doi.org/10.1016/j.atmosenv.2021.118649.
11. Sahney, Manju, and Swati Chaurasia. "SEASONAL VARIATIONS OF AIRBORNE POLLEN IN ALLAHABAD, INDIA." *Annals of Agricultural and*

View publication stats

Environmental Medicine, vol. 15, no. 2, Dec. 2008, pp. 287–93. *www.aaem.pl*, https://www.aaem.pl/SEASONAL-VARIATIONS-OF-AIRBORNE-POLLEN-IN-ALLAHABAD-INDIA,90515,0,2.html.

12. Ramon, G. D., Vanegas, E., Felix, M., Barrionuevo, L. B., Kahn, A. M., Bertone, M., Reyes, M. S., Gaviot, S., Ottaviano, C., & Cherrez-Ojeda, I. (2020). Year-long trends of airborne pollen in Argentina: More research is needed. *The World Allergy Organization journal*, *13*(7), 100135. https://doi.org/10.1016/j.waojou.2020.100135
13. Frenz, D. A., & Lince, N. L. (1997). A comparison of pollen recovery by three models of the Rotorod sampler. *Annals of allergy, asthma & immunology : official publication of the American College of Allergy, Asthma, & Immunology*, *79*(3), 256–258. https://doi.org/10.1016/S1081-1206(10)63011-6
14. Levetin, Estelle, et al. "Comparison of Pollen Sampling with a Burkard Spore Trap and a Tauber Trap in a Warm Temperate Climate." *Grana*, vol. 39, no. 6, Jan. 2000, pp. 294–302. *DOI.org (Crossref)*, https://doi.org/10.1080/00173130052504333.
15. Papadogiannaki S, Kontos S, Parliari D, Melas D. Machine Learning Regression to Predict Pollen Concentrations of Oleaceae and Quercus Taxa in Thessaloniki, Greece. *Environmental Sciences Proceedings*. 2023; 26(1):2. https://doi.org/10.3390/environsciproc2023026002
16. Buters, J. T. M., Antunes, C., Galveias, A., Bergmann, K. C., Thibaudon, M., Galán, C., Schmidt-Weber, C., & Oteros, J. (2018). Pollen and spore monitoring in the world. *Clinical and translational allergy*, *8*, 9. https://doi.org/10.1186/s13601-018-0197-8
17. Picornell, A., et al. "Increasing Resolution of Airborne Pollen Forecasting at a Discrete Sampled Area in the Southwest Mediterranean Basin." Chemosphere, vol. 234, Nov. 2019, pp. 668–81. ScienceDirect, https://doi.org/10.1016/j.chemosphere.2019.06.019
18. Mills, Sophie A., et al. "Machine Learning Methods for Low-Cost Pollen Monitoring – Model Optimisation and Interpretability." Science of The Total Environment, vol. 903, Dec. 2023, p. 165853. ScienceDirect, https://doi.org/10.1016/j.scitotenv.2023.165853