Spatio-Temporal Analysis between Mining Production and ARI Incidence in Arequipa: A Data Engineering and Environmental Epidemiology Approach

Equipo Bumblebees
School of Systems Engineering
Universidad Nacional de San Agustín de Arequipa (UNSA)
Arequipa, Peru
Email: ycozco@unsa.edu.pe

***Abstract*—The Arequipa region faces ongoing tension between its role as the mining hub of southern Peru and public concern over potential impacts on respiratory health. Although the State produces abundant open data—mining bulletins, epidemiological reports, and demographic projections—these had not been systematically integrated to evaluate the relationship between mining production and the incidence of Acute Respiratory Infections (ARI, *IRA* in Spanish).**

This study develops a spatio–temporal analysis for the 2021–2023 period, based on an ETL *pipeline* implemented in Python that unifies information from MINEM, GERESA, and INEI. The processing resolves temporal discrepancies, district naming differences, and population variations through incidence rates normalized per 10 000 inhabitants.

The scientific objective was to test the null hypothesis of no immediate linear correlation between monthly mining production and ARI rates. The results show correlation coefficients close to zero at the regional and provincial levels, indicating that production peaks do not generate simultaneous increases in respiratory morbidity. However, the comparative analysis between districts reveals a baseline gap: districts with mining activity exhibit slightly higher ARI rates, even while sharing the same climatic seasonality.

The work provides a reproducible methodological framework to audit relationships between extractive activity and public health using only open data, establishing a quantitative basis for future research that integrates environmental variables and more complex spatio–temporal models.

***Index Terms*—Mining, Acute Respiratory Infections, Data Engineering, ETL, Environmental Epidemiology, Arequipa.**

I. Introduction

The Arequipa region occupies a central place in the economic structure of Peru: it concentrates large-scale mining operations, generates a significant share of the mining canon, and hosts strategic infrastructure for the export of metallic and non-metallic minerals. At the same time, reports from the Regional Health Office show a high burden of Acute Respiratory Infections (ARI), especially in children and older adults during the winter months.

In this context, a recurring question emerges in the media, community assemblies, and political spaces: *"Is mining making the population sick?"*. The strength of this question stems not only from statistics but also from the everyday experience of communities that live with dust in suspension, high-tonnage trucks, and constant industrial emissions. However, public debate has developed largely without an integrated analysis of the available data.

Currently, State institutions generate valuable but fragmented information. MINEM publishes monthly production by mineral, mining unit, and district; GERESA reports weekly cases of ARI, pneumonia, and associated deaths; INEI provides district-level population projections that allow normalized rates to be calculated. Each dataset, on its own, describes one aspect of reality. What is missing is a technical bridge that allows them to be crossed coherently to obtain a *"complete picture"* of the mining–health relationship.

This work arises precisely from that need. In the context of the Software Project Management course, a Data Engineering *pipeline* was designed and implemented to articulate the extraction, cleaning, transformation, and loading (*Extract–Transform–Load, ETL*) of data from MINEM, GERESA, and INEI for the 2021–2023 period. Based on this *pipeline*, an integrated dataset was built that makes it possible to analyze, at the district and provincial levels, the potential relationship between mining production and ARI incidence. To ensure transparency and reproducibility, the source code and processed data have been published in an open repository (https://github.com/ycozco/HADS_Bumblebees).

The main scientific objective is to test the following null hypothesis:

$$H_0 : \rho(\text{Monthly Mining Production}, \text{ARI Rate}) \approx 0 \qquad (1)$$

where ρ denotes the linear correlation coefficient between monthly mining production, expressed in fine metric tons (TMF), and the incidence rate of ARI normalized per 10 000 inhabitants. Rejecting this hypothesis would imply finding statistical evidence of a direct and immediate association; failing to reject it would suggest that morbidity peaks are driven by other dominant factors, such as climatic seasonality, even in the presence of intense mining activity.

Beyond global correlation, the study seeks to answer two additional questions:

- **P1.** Do districts with mining activity show, on average, higher ARI rates than districts without mining activity, even when they share the same climatic seasonality?
- **P2.** Does the production–ARI relationship exhibit temporal lags that suggest cumulative or delayed effects over time?

Answering these questions requires not only statistical tools but also special care in how the dataset is constructed and validated. For this reason, the focus of the article is not limited to numerical results but also includes a detailed description of the engineering decisions adopted in the ETL *pipeline*.

II. Justification of the Research

The central motivation of this work can be summarized in a simple idea: in Arequipa there is sufficient public information to rigorously analyze the mining–health relationship, but this information has not yet been articulated into an integrated and accessible analytical tool. What follows is, in essence, a justification of why it was worthwhile to invest effort in building this tool and documenting the process.

A. Social and Territorial Justification

Arequipa combines a strong economic dependence on mining with a history of socio-environmental conflicts linked to perceived health risks. In practice, this translates into a daily duality: on the one hand, the population recognizes the benefits of the mining canon, employment, and investment; on the other hand, it expresses concern about dust in suspension, noise, heavy traffic, and possible air and water pollution.

In this context, public debate tends to rely on testimonies or isolated cases ("my neighbor got sick after the mine expanded"), but rarely on comparable time series or reproducible statistical analyses. The absence of integrated evidence fuels polarization: each actor selects the data that reinforce their position, without a shared frame of reference.

This study seeks to reduce that gap by offering an analysis based exclusively on official data from the Peruvian State itself. The choice of public sources is not accidental: it aims to minimize accusations of bias, both for and against mining activity, and to demonstrate that the same information already being reported can be better leveraged using data engineering tools.

B. Scientific Justification

From an academic perspective, the research is justified because it addresses a question that combines three traditionally separate domains: mining production, respiratory epidemiology, and spatio-temporal data analysis. Previous studies have demonstrated the relationship between particulate matter pollution and respiratory diseases in urban or industrial contexts; however, there are fewer works that explicitly cross officially reported mining production with health indicators at the district level.

The project proposes a methodological framework that could be replicated in other regions of the country or in other extractive contexts, as long as sufficient open data exist. In this way, the contribution is not limited to the Arequipa case but extends as an example of how Data Engineering can support decision-making in public health.

C. Value Proposition: From Perception to Evidence

The value proposition of the work can be stated directly: *moving from subjective perception to quantitative evidence*. This does not mean negating the validity of local experience, but rather complementing it with indicators that allow hypotheses to be tested against data.

The study adopts a deliberately neutral stance: instead of assuming that mining necessarily makes people sick or, conversely, that it has no impact at all, it sets out an explicit null hypothesis and designs a set of analyses to attempt to refute it. If the evidence is not sufficient to reject it, the interpretation will differ from that which would arise if clear patterns of association were found.

D. Scientific Objective and Guiding Question

The main scientific objective is to evaluate the presence or absence of an immediate linear correlation between monthly mining production and ARI incidence, along with characterizing possible differential effects between mining and non-mining districts.

In simple terms, the project's guiding question can be expressed as follows: *"If the tons extracted increase in a specific month, is there a proportional increase in respiratory patients in that same period or in the following months, and is this effect different in mining versus non-mining areas?"*.

Answering this question with real data, even if the final answer is "there is no immediate correlation", provides clarity to public debate and better guides future efforts in environmental and epidemiological monitoring.

III. Justification of Processes and Engineering Decisions

In addition to the social and scientific motivation, the project is justified by the technical value of the developed *pipeline*. This section explains the key design decisions, which are non-trivial and directly condition the validity of the results.

A. Temporal Unification: Choosing the Month as Unit of Analysis

The first technical conflict encountered was the difference in temporal granularity between sources: GERESA reports ARI cases by Epidemiological Week (from 1 to 52 or 53), whereas MINEM publishes mining production on a monthly basis. Forcing mining data into a weekly scale would require arbitrary assumptions about the intra-monthly distribution of production; by contrast, aggregating epidemiological weeks into calendar months is a more conservative operation from a statistical standpoint.

Therefore, the **month** was defined as the *minimum common temporal denominator*. Operationally, a week–month map was constructed that assigns each Epidemiological Week to the month in which most of its days fall. In this way, health data are aggregated upward (from week to month), while production remains at its original resolution.

This decision has direct implications: the study explicitly renounces analyzing very short-term phenomena (for example, outbreaks lasting a few days) in order to gain consistency in the mining–health comparison. The objective is not to model each specific outbreak but to identify patterns month by month.

B. Population Normalization: From Absolute Numbers to Rates

A frequent error in exploratory epidemiological analyses is comparing absolute numbers of cases between districts with very different population sizes. In Arequipa, districts such as Cerro Colorado or Paucarpata concentrate tens of thousands of inhabitants, while rural areas such as Pocsi or San Juan de Tarucani have much smaller populations. Under such conditions, absolute ARI cases tend to be higher in urban districts simply because more people are exposed, not necessarily because individual risk is higher.

To avoid this bias, the **Normalized ARI Incidence Rate** was defined as:

$$\text{Tasa}_{ARI} = \left(\frac{\text{Total Cases in the Month}}{\text{Projected District Population}} \right) \times 10\,000 \quad (2)$$

The scaling factor 10 000 was chosen for readability: it allows rates to take numerically manageable values that are comparable between districts without resorting to scientific notation. This metric converts raw case counts into a relative risk indicator that can indeed be compared between large and small districts.

C. Resolution of "Orphan Districts": Entity Cleaning and Matching

During the initial merging of the databases, it was found that approximately 30% of health records did not match any record in the mining database, and vice versa. The issue was not the absence of activity but discrepancies in district naming: GERESA might record "DISTRITO DE YURA" or "YURA (CS)", while MINEM simply used "YURA"; in other cases, accents, abbreviations, or capitalization differed.

Ignoring these cases would have introduced a severe bias, particularly because several of the affected districts correspond precisely to areas of mining interest (Yura, Uchumayo, San Juan de Tarucani). To address this, an *entity resolution* process was implemented based on:

- Text normalization to uppercase.
- Removal of accents using the `unidecode` library.
- Cleaning of administrative prefixes and suffixes ("DISTRITO DE", "(CS)", etc.).
- Manual verification of critical cases.

The result was the recovery of traceability for practically all relevant districts. From a data quality perspective, this step was crucial: without it, the analyses would have artificially suggested that in certain mining districts "there are no patients", when in fact the problem was purely one of nomenclature.

D. Handling Missing Values: NaN versus Zero

Another key engineering decision was the treatment of missing values in the reports. The following policy was adopted:

- **Missing mining production** → interpreted as zero production, assuming that the absence of a report in a specific period implies no recorded activity or that the project's scale does not significantly affect regional volume.
- **Missing health reports** → kept as `NaN` (not available), avoiding arbitrary replacement with zero.

This distinction follows a principle of epidemiological prudence: recording "zero patients" when a health center simply did not submit information could artificially lower the average morbidity rate and mask potential underreporting issues. Leaving the value as `NaN` is statistically more honest, even if it means that certain months or districts are excluded from some aggregate calculations.

IV. Theoretical Framework

The empirical analysis of the mining–ARI relationship requires a conceptual framework that brings together three domains: the epidemiology of respiratory infections, the dynamics of atmospheric pollution in mining contexts, and the basic principles of spatio-temporal analysis of health data. This section presents a synthesis of these elements, with emphasis on aspects that are directly relevant to the Arequipa region.

A. Acute Respiratory Infections and Vulnerability in Highland Areas

Acute Respiratory Infections are one of the main reasons for medical consultation and hospitalization in Peru, particularly among children and older adults. The epidemiological literature highlights that ARI incidence is strongly modulated by environmental and social factors: temperature, humidity, overcrowding, indoor air quality, and access to health services, among others.

In highland areas such as Arequipa, the seasonal pattern of ARI shows a marked increase during the cold months. The combination of low nighttime temperatures, dwellings with limited thermal insulation, and exposure to biomass combustion indoors increases susceptibility to infections of the upper and lower respiratory tract. This seasonal pattern has been documented in Ministry of Health reports and constitutes a fundamental starting point: any analysis of the potential role of mining must acknowledge that, even in the absence of extractive activity, climate alone generates morbidity peaks.

B. Particulate Matter Pollution and Respiratory Health

Various international studies have demonstrated the association between elevated concentrations of respirable particulate matter (PM_{10} and $PM_{2.5}$) and higher rates of respiratory diseases, asthma exacerbations, hospitalizations, and premature mortality. The underlying mechanism includes irritation of the airways, chronic inflammation, and, in the case of particles containing heavy metals, oxidative stress.

In mining contexts, sources of particulate matter include rock extraction and crushing, hauling of ore in high-tonnage trucks, blasting, and, in some cases, smelting processes. Particle dispersion also depends on wind direction and intensity, precipitation, and local topography. The most reasonable working hypothesis, therefore, is not that mining causes a sudden spike in ARI cases in a specific month, but rather that it contributes to an increase in the *baseline risk* of the exposed population, making seasonal viral episodes more severe.

C. Seasonality, Trends, and Lags

Time-series analysis in epidemiology distinguishes three main components: trend, seasonality, and noise. The trend captures long-term changes; seasonality reflects recurring patterns in specific periods of the year; noise encompasses unexplained variations. In the Arequipa region, ARI data exhibit clear seasonality linked to winter, which introduces a methodological challenge: if this seasonality is not controlled for, any comparison between districts can be confounded by the simple fact that they share the same regional climate.

Furthermore, certain environmental factors, such as air pollution, may not manifest their effects instantaneously but rather with a temporal lag. Thus, the relationship between mining production and health may not be strictly simultaneous (same month), but may involve cumulative or delayed effects. In practice, the availability of monthly data limits the capacity to detect fine-grained lags, but the theoretical framework invites us to consider this possibility when interpreting results.

D. Spatio-Temporal Analysis and District-Level Comparison

The spatial component of the analysis is equally important. It is not enough to evaluate a global correlation between regional production and total ARI: it is necessary to distinguish between districts with direct presence of mining units and districts without extractive activity, and to compare their epidemiological trajectories over time. This approach makes it possible to assess whether, beyond shared seasonality, there is a systematic difference in baseline morbidity between the two groups.

The literature on spatial epidemiology frequently uses techniques such as Moran's autocorrelation index, geographically weighted regression (GWR), or spatial autoregressive models (SAR/SEM) to capture dependency patterns between territorial units. Although these methods exceed the scope of the current phase of the project, they provide the theoretical basis for the future lines of work that will be proposed in later sections.

V. Data Methodology and Statistical Analysis Design

Once the conceptual framework was established, the next step was to translate the research questions into a concrete analysis design. This section details the construction of the integrated dataset, the definition of variables, the levels of aggregation considered, and the statistical methods used to test the null hypothesis and explore spatio–temporal patterns.

A. Construction of the Integrated Dataset

The ETL *pipeline* developed in Python made it possible to obtain, for each *district–month–year* combination, the following minimum information:

- Total monthly mining production (in TMF), aggregated by district.
- ARI, pneumonia, and respiratory death cases, summed by month.
- Projected district population for the corresponding year.
- Binary label indicating whether the district has relevant mining activity.

In the case of provinces, the information was obtained by aggregating (by sum for production and health cases, or by weighted sum for incidence rate) all districts contained in each province.

B. Study Variables

Table I summarizes the set of variables analyzed. Note that the emphasis is placed on the normalized incidence rate rather than on absolute counts.

TABLE I
Main variables used in the analysis

Variable	Description
Producción_TMF	Total monthly mining production of the district (TMF), sum of metals and non-metals reported by MINEM.
Casos_ARI	Monthly total of ARI cases reported by GERESA (including ARI without pneumonia, pneumonias, and associated deaths).
Población	District population projected by INEI for the corresponding year.
Tasa_ARI	Casos_ARI normalized by population and scaled per 10 000 inhabitants.
Mes	Month identifier (1–12), used to capture seasonality.
Distrito_Minero	Binary indicator (1 = district with reported mining activity; 0 = district without mining activity).

C. Levels of Analysis

Three complementary levels were defined:

1) **Global level**: the entire Arequipa region considered as a single system; correlation between total production and total incidence is evaluated.
2) **Provincial level**: comparison between provinces, identifying differentiated patterns according to their degree of mining activity.
3) **District level**: detailed analysis of selected districts (Yura, Uchumayo, Callalli, San Juan de Tarucani, among others), with emphasis on local dynamics of production and disease.

Each level offers a different perspective: global analysis assesses the strength of the average signal; provincial analysis reveals intermediate heterogeneities; district-level analysis approaches the scale at which environmental and health impacts are experienced.

D. Statistical Procedures

To test the null hypothesis and characterize the mining–ARI relationship, the following procedures were applied:

1) Pearson Correlation: The Pearson correlation coefficient was used to evaluate the existence of a linear relationship between monthly production (in logarithmic scale) and ARI rate. The logarithmic transformation of production was used to reduce skewness in the TMF distribution, which is typical of economic data with highly variable magnitudes.

2) Spearman Correlation: Since normality of the variables cannot be guaranteed, Spearman's correlation was additionally calculated. This measure is based on ranks and allows exploration of monotonic, not necessarily linear, relationships. Consistency between both coefficients reinforces interpretation.

3) Comparison of Means between Mining and Non-Mining Districts: To assess whether districts with mining activity exhibit a different baseline ARI incidence than districts without mining activity, a difference-in-means test was applied. Given that sample sizes and variances may differ between groups, the robust version of Student's *t*-test with Welch's correction was adopted. The contrast was performed both on the average monthly rate and on indicators aggregated by year.

4) Analysis of Seasonality and Lags: Seasonality was explored visually using time series and trend plots by month. In addition, possible lags between production and ARI rate were evaluated through cross-correlation analysis between the production series and the rate series, considering lags of one to three months. Although monthly resolution limits the power to detect rapid effects, this analysis allows the identification of first-order delayed patterns.

5) Construction of the "Offset" Indicator: To formalize the observed gap between mining and non-mining districts, an *offset* indicator was built, defined as the average monthly difference between rates of both groups, with seasonality fixed. This indicator is interpreted as an approximation of the *additional baseline risk* associated with residing in a mining district, independent of climatic patterns.

VI. Results

This section presents the main findings of the quantitative analysis. The structure of levels defined earlier is followed: first, the global relationship is discussed, then provincial patterns, and finally district-level behavior and the mining–non-mining comparison.

A. Global Production–ARI Relationship

Figure 1 shows the scatter plot between total monthly mining production (in logarithmic scale) and the corresponding regional ARI rate for the same month.

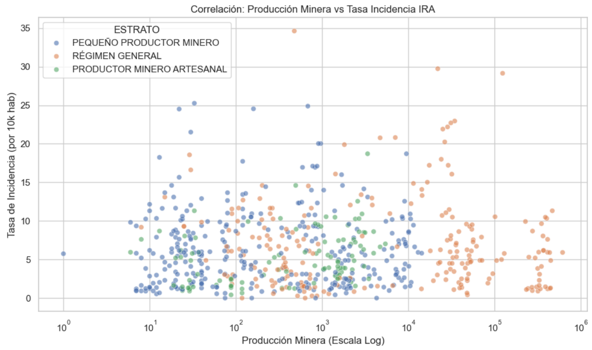

Fig. 1. Scatter plot between monthly mining production (log-TMF) and ARI incidence rate in Arequipa.

The cloud of points is clearly diffuse, with no evident upward or downward trend. Pearson and Spearman correlation coefficients computed on this dataset are very close to zero, indicating the absence of a strong linear or monotonic relationship between the variables. In terms of the null hypothesis stated in the Introduction, this result suggests that, at least at the regional level and on a monthly scale, no immediate effect of mining production on ARI incidence is observed.

B. Provincial Behavior

To explore possible intermediate heterogeneities, trend plots were constructed at the provincial level. Figure 2 presents the evolution of average ARI rates by province, while Figures 3–7 show, for each selected province, the joint series of mining production and ARI rate.

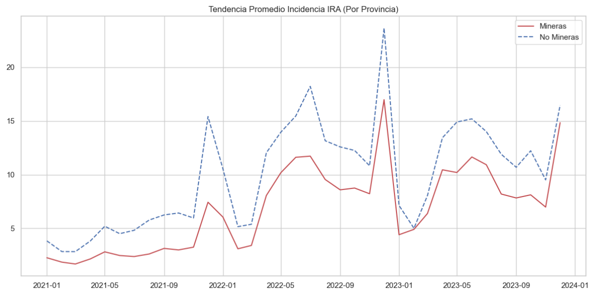

Fig. 2. Temporal trend of ARI rates at the provincial level (Arequipa, Camaná, Caravelí, Caylloma, and Islay).

In all provinces, a very similar seasonal pattern is observed: rates increase during autumn and winter and decrease in spring–summer. This provincial synchronicity supports the hypothesis that climate is the main driver of ARI peaks, over and above specific local factors.

Figures 3 to 7 superimpose provincial mining production (bars or line) and ARI rate (line), allowing the local relationship between the two series to be visualized.

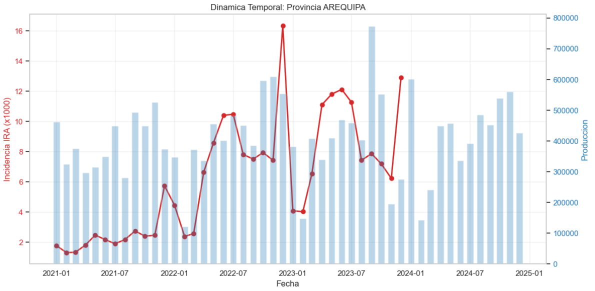

Fig. 3. Arequipa Province: monthly mining production vs. ARI rate.

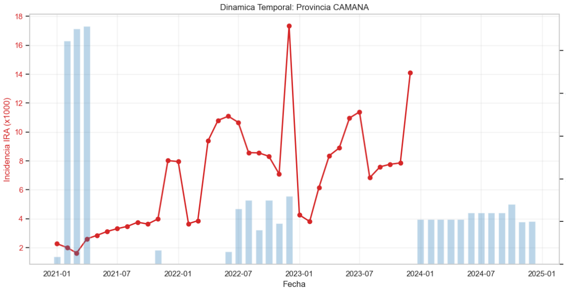

Fig. 4. Camaná Province: monthly mining production vs. ARI rate.

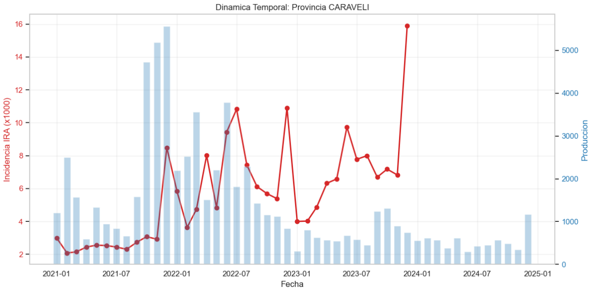

Fig. 5. Caravelí Province: monthly mining production vs. ARI rate.

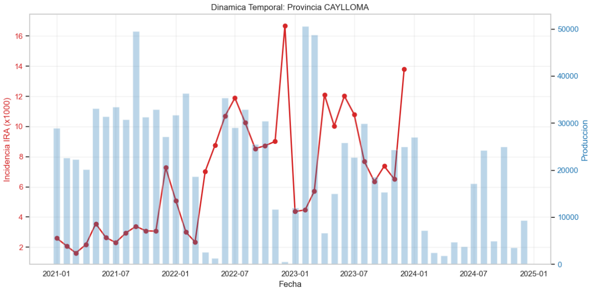

Fig. 6. Caylloma Province: monthly mining production vs. ARI rate.

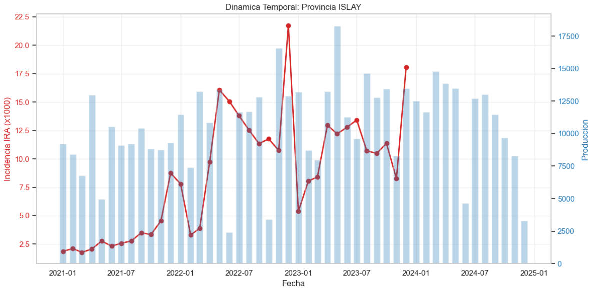

Fig. 7. Islay Province: monthly mining production vs. ARI rate.

In none of the cases is a consistent direct relationship observed between production peaks and disease peaks. Instead, the dominant feature is again the seasonal cycle in the epidemiological curve, relatively independent from production fluctuations.

C. District-Level Analysis: Yura, Uchumayo, Callalli, and San Juan de Tarucani

District-level analysis zooms in on territorial units directly exposed to mining operations. Figures 8 to 11 show dual-axis plots for four emblematic districts.

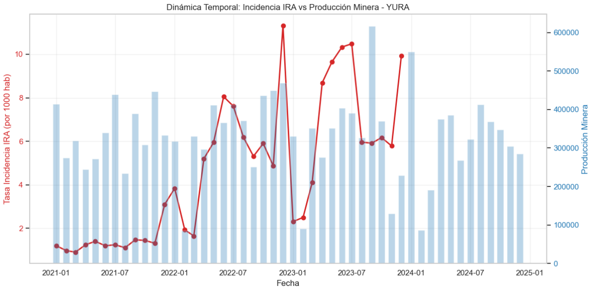

Fig. 8. Yura District: monthly mining production vs. ARI rate.

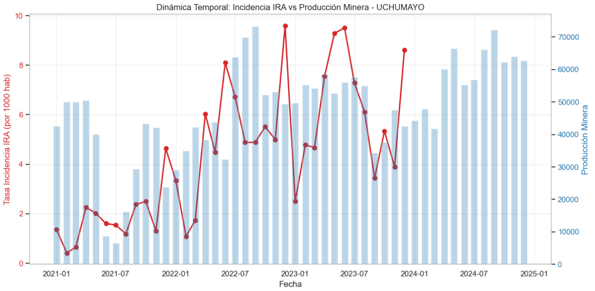

Fig. 9. Uchumayo District: monthly mining production vs. ARI rate.

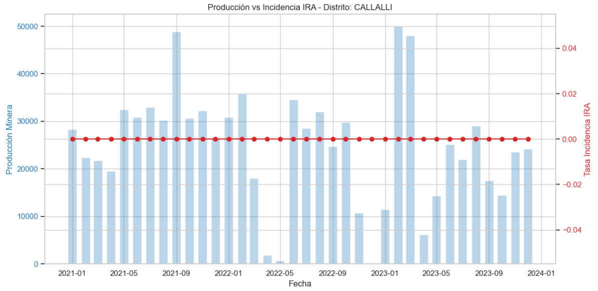

Fig. 10. Callalli District: monthly mining production vs. ARI rate.

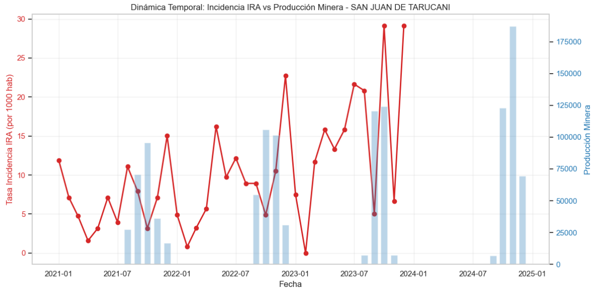

Fig. 11. San Juan de Tarucani District: monthly mining production vs. ARI rate.

In all four districts, the same phenomenon is observed: the ARI rate curve follows an annual cycle similar to that of the rest of the region, while mining production displays peaks and troughs that do not systematically coincide with disease peaks. This pattern of *asynchrony* supports the conclusion that monthly extractive activity does not, on its own, explain acute ARI outbreaks.

D. Comparison between Mining and Non-Mining Districts and the "Offset" Phenomenon

To assess structural differences between mining and non-mining districts, two cohorts were constructed: one grouping districts with mining activity and another grouping districts without such activity. Figure 12 shows a boxplot comparing the distribution of ARI rates between both groups, while Figure 13 presents the temporal evolution of their averages.

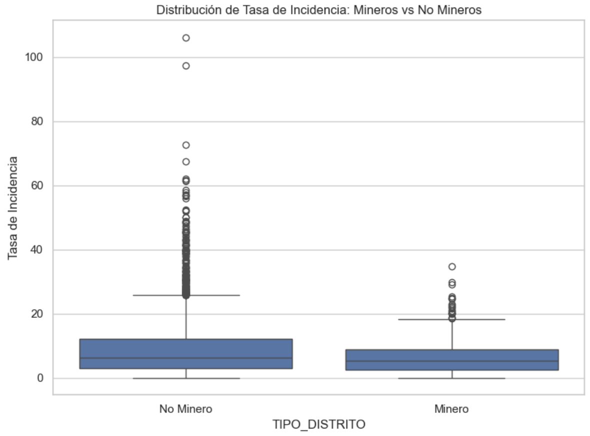

Fig. 12. Comparative distribution of ARI rates in mining vs. non-mining districts.

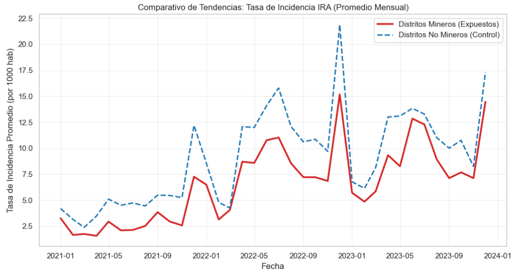

Fig. 13. Temporal trends of ARI rates in mining and non-mining districts.

The boxplot suggests that the median and a large part of the rate distribution in mining districts lie above those of non-mining districts. The trend curves show that both series share the same seasonality (rising and falling at the same times), but the mining-district series remains consistently a few points higher: this is the identified *offset* or baseline gap.

The difference-in-means test, applied to monthly rates averaged by group, indicates that this difference is statistically significant under a reasonable confidence level. Although the magnitude of the gap is not sufficient to attribute direct causality to mining, it does suggest that populations in mining districts face a slightly higher respiratory risk throughout the year.

VII. Discussion

The results make it possible to answer, in a nuanced way, the question that motivated the project. First, global correlation analyses and scatter plots show that there is no strong linear association between monthly mining production and ARI rates in the Arequipa region. In terms of the stated null hypothesis, the empirical evidence supports the conclusion that monthly production peaks do not, by themselves, generate simultaneous peaks in respiratory disease.

Second, provincial and district analyses confirm that the dominant component in ARI dynamics is climatic seasonality.

Both mining and non-mining districts experience more intense outbreaks during the cold months, consistent with the epidemiological literature on highland areas. This finding is important because it contextualizes social perceptions: the coincidence of winter with certain mining activities can lead people to interpret as causality what is, to a large extent, an effect of climate.

However, the research does not merely conclude that "mining has no influence"; in fact, the results reveal a subtler pattern. The comparison between mining and non-mining districts shows a systematic difference in baseline ARI rates: even when both series share the same seasonal cycle, the curve for mining districts is consistently higher. This gap, although moderate, is robust and statistically significant.

From an epidemiological perspective, this *offset* can be interpreted as an increase in chronic vulnerability among populations residing in mining areas. Possible mechanisms that could explain this situation include:

- Sustained exposure to ambient particulate matter, resulting from extractive activity and heavy-vehicle traffic.
- Specific socioeconomic conditions in mining areas (precarious housing, unequal access to health services, higher levels of occupational stress).
- Interactions between ambient dust and pre-existing risk factors (wood-smoke exposure, extreme climatic conditions).

The study does not have enough data to discriminate among these explanations, but it does establish a solid starting point: mining does not appear to be the immediate driver of seasonal peaks, but it may be related to a sustained elevation of respiratory risk.

Finally, the methodological value should be highlighted. Beyond the specific results for Arequipa, the data-integration *pipeline* shows that it is possible to audit the mining–health relationship using only open data and free software tools. This opens the door to replicating the approach in other regions of the country or in other productive sectors.

VIII. Study Limitations

Like any observational study based on secondary data, this work has several limitations that must be explicitly acknowledged:

- **Temporal resolution**: choosing the month as the unit of analysis, although methodologically coherent, prevents detection of short-term effects (for example, particle spikes on specific days).
- **Quality of health data**: the ARI rate depends on the timeliness and completeness of reporting by health facilities. Underreporting is possible, especially in remote rural areas.
- **Absence of direct environmental data**: complete time series of PM_{10}, $PM_{2.5}$, temperature, and humidity were not available for the analyzed districts. This limits the ability to explicitly model the role of air pollution.
- **Causality vs. correlation**: the study's approach is correlational; even if strong associations were observed, causality could not be attributed without a more complex experimental or quasi-experimental design.
- **Internal heterogeneity of mining activity**: the analysis considers total production per district, without differentiating between mineral types, technologies used, or environmental mitigation practices, which may have very different impacts.

These limitations do not invalidate the results but do circumscribe their scope and highlight the need to complement this type of analysis with additional information.

IX. Future Work

Based on the experience gained in this project, several lines of future work are identified:

- **Integration of environmental data**: incorporating historical series of PM_{10} and $PM_{2.5}$ concentrations, as well as meteorological variables (temperature, humidity, wind speed and direction), would enable more robust multivariate models.
- **Advanced spatial analysis**: applying spatial autocorrelation statistics (such as Moran's index) and spatial regression models (SAR, SEM, GWR) to capture dependence among neighboring districts and assess geographical propagation of risk.
- **Development of an interactive platform**: packaging the ETL *pipeline* and results into an interactive dashboard accessible to authorities and citizens, fostering transparency and informed participation.
- **Predictive models**: exploring machine learning techniques (for example, random forests or recurrent neural networks) to predict ARI burden from climatic, productive, and environmental variables, evaluating different intervention scenarios.
- **Linking with clinical and community studies**: complementing quantitative analysis with surveys, clinical measurements, and household-level air quality monitoring in selected districts, in order to validate hypotheses about exposure and vulnerability mechanisms.

Taken together, these lines of work would allow a transition from the initial quantitative diagnosis to a deeper and more actionable understanding of the mining–health relationship in Arequipa.

X. Conclusions

This study developed and applied a Data Engineering *pipeline* to integrate mining, epidemiological, and demographic information for the Arequipa region during the 2021–2023 period. Based on this unified dataset, the possible relationship between monthly mining production and the incidence of Acute Respiratory Infections was evaluated, considering different levels of analysis (regional, provincial, and district).

The main findings can be summarized as follows:

1) At the regional and provincial levels, correlation coefficients between monthly mining production and ARI rate are close to zero, supporting the null hypothesis of no immediate linear relationship between the two variables.

2) Climatic seasonality emerges as the dominant factor in ARI dynamics: all districts, mining and non-mining, show morbidity peaks during the cold months, regardless of production fluctuations.
3) Even though seasonal peaks are shared, districts with mining activity exhibit, on average, slightly higher ARI rates than non-mining districts, forming an *offset* or baseline gap that suggests possible chronic vulnerability associated with the mining context.
4) The developed ETL *pipeline* demonstrates that it is possible to audit the mining–health relationship using only open data and free software tools, representing a methodological contribution that can be replicated in other regions and contexts.

Taken together, the results invite us to abandon simplistic explanations. Mining does not appear to be the direct trigger of acute ARI outbreaks, but it may be contributing to a sustained increase in respiratory risk among exposed populations. Confirming or refining this hypothesis will require incorporating environmental data and more complex research designs, but the present work establishes a quantitative foundation on which such efforts can build.

Acknowledgements

The authors thank the teaching team of the School of Systems Engineering at UNSA for their methodological guidance during the development of the project, as well as the institutions that produce open data (MINEM, GERESA, INEI) whose transparency efforts made this study possible.

References

[1] Ministerio de Energía y Minas del Perú (MINEM), “Mining Statistical Bulletin 2021–2023,” Lima, Peru.

[2] Gerencia Regional de Salud de Arequipa (GERESA), “Reports from the Acute Respiratory Infections Situational Room,” 2021–2023.

[3] Instituto Nacional de Estadística e Informática (INEI), “District Population Projections 2018–2025,” Lima, Peru.

[4] World Health Organization (WHO), “Air Pollution and Child Health: Prescribing Clean Air,” 2018.

[5] Centers for Disease Control and Prevention (CDC), “Seasonal Patterns of Respiratory Infections in High-Altitude Regions,” Technical Report, 2020.

[6] World Health Organization, “Ambient (Outdoor) Air Pollution,” Fact Sheet, 2021.

[7] J. Samet and D. Krewski, “Health Effects Associated with Exposure to Ambient Air Pollution,” *Journal of Toxicology and Environmental Health*, vol. 68, pp. 1–28, 2014.

[8] A. Lawson, *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology*, 3rd ed., CRC Press, 2018.

[9] Equipo Bumblebees, “HADS Project Repository: Mining-Health Analysis in Arequipa,” GitHub, 2025. [Online]. Available: https://github.com/ycozco/HADS_Bumblebees.