Análisis Espacio-Temporal entre Producción Minera e Incidencia de IRA en Arequipa: Un Enfoque de Ingeniería de Datos y Epidemiología Ambiental

Equipo Bumblebees
Escuela Profesional de Ingeniería de Sistemas
Universidad Nacional de San Agustín de Arequipa (UNSA)
Arequipa, Perú
Email: hquispeh@unsa.edu.pe

***Abstract*—La región Arequipa vive una tensión permanente entre su rol como núcleo minero del sur del Perú y la percepción ciudadana de que esta actividad deteriora la salud respiratoria de la población. A pesar de la abundancia de datos abiertos producidos por el Estado –boletines mineros, reportes epidemiológicos y proyecciones demográficas–, hasta ahora no se contaba con un estudio que integrara de manera sistemática estas fuentes para evaluar, con evidencia cuantitativa, la relación entre producción minera e incidencia de Infecciones Respiratorias Agudas (IRA).**

Este trabajo presenta un análisis espacio-temporal desarrollado a partir de un *pipeline* ETL implementado en Python que consolida información del Ministerio de Energía y Minas (MINEM), la Gerencia Regional de Salud (GERESA) y el Instituto Nacional de Estadística e Informática (INEI) para el periodo 2021–2023. Se resuelven incompatibilidades temporales (semanas epidemiológicas frente a meses calendario), inconsistencias de nomenclatura distrital y diferencias de escala poblacional mediante tasas de incidencia normalizadas por cada 10 000 habitantes.

El objetivo científico central es contrastar la hipótesis nula de que la producción minera mensual no presenta correlación lineal inmediata con la tasa de IRA. Adicionalmente, se busca identificar si los distritos con actividad minera presentan un nivel basal de morbilidad diferente al de distritos sin actividad extractiva, controlando por la estacionalidad climática. El artículo combina elementos de ingeniería de datos, estadística aplicada y epidemiología ambiental, proponiendo un marco metodológico reproducible para futuros estudios sobre minería y salud pública en regiones andinas.

***Index Terms*—Minería, Infecciones Respiratorias Agudas, Ingeniería de Datos, ETL, Epidemiología Ambiental, Arequipa.**

I. Introducción

La región Arequipa ocupa un lugar central en la estructura económica del Perú: concentra operaciones mineras de gran escala, genera una fracción significativa del canon y alberga infraestructura estratégica para la exportación de minerales metálicos y no metálicos. Al mismo tiempo, los reportes de la Gerencia Regional de Salud evidencian una alta carga de Infecciones Respiratorias Agudas (IRA), especialmente en población infantil y adulta mayor durante los meses de invierno.

En este contexto emerge una pregunta que se repite en medios de comunicación, asambleas vecinales y espacios políticos: *"¿la minería está enfermando a la población?"*. La fuerza de esta pregunta no proviene únicamente de la estadística, sino de la experiencia cotidiana de las comunidades que conviven con polvo en suspensión, camiones de alto tonelaje y emisiones industriales constantes. Sin embargo, la discusión pública se ha desarrollado, en gran medida, sin un análisis integrado de los datos disponibles.

Actualmente, las instituciones del Estado generan información valiosa pero fragmentada. El MINEM publica mensualmente la producción por mineral, unidad minera y distrito; la GERESA reporta semanalmente los casos de IRA, neumonía y defunciones asociadas; el INEI provee proyecciones poblacionales distritales que permiten calcular tasas normalizadas. Cada conjunto de datos, por separado, describe un aspecto de la realidad. Lo que falta es un puente técnico que permita cruzarlos de manera coherente para obtener una *"fotografía completa"* de la relación minería–salud.

El presente trabajo surge precisamente de esa necesidad. En el marco del curso Gestion de Proyectos de Software, se diseñó e implementó un *pipeline* de Ingeniería de Datos que articula la extracción, limpieza, transformación y carga (*Extract–Transform–Load, ETL*) de los datos de MINEM, GERESA e INEI para el periodo 2021–2023. A partir de este *pipeline*, se construye un dataset integrado que permite analizar, a nivel distrital y provincial, la posible relación entre producción minera e incidencia de IRA. Para garantizar la transparencia y reproducibilidad del estudio, el código fuente y los datos procesados se han publicado en un repositorio abierto (https://github.com/ycozco/HADS_Bumblebees).

El objetivo científico principal es contrastar la siguiente hipótesis nula:

$$H_0 : \rho(\text{Producción Minera Mensual}, \text{Tasa de IRA}) \approx 0 \quad (1)$$

donde ρ denota el coeficiente de correlación lineal entre la producción minera mensual, expresada en toneladas métricas finas (TMF), y la tasa de incidencia de IRA normalizada por cada 10 000 habitantes. Rechazar esta hipótesis implicaría encontrar evidencia estadística de una asociación directa e inmediata; no rechazarla sugeriría que los picos de morbilidad obedecen a otros factores dominantes, como la estacionalidad climática, aun en presencia de actividad minera intensa.

Más allá de la correlación global, el estudio busca responder dos preguntas adicionales:

- **P1.** ¿Los distritos con actividad minera presentan, en promedio, una tasa de IRA mayor que los distritos sin actividad minera, aun cuando comparten la misma estacionalidad climática?
- **P2.** ¿La relación producción–IRA presenta retardos temporales (desfasajes) que sugieran efectos acumulativos o diferidos en el tiempo?

Responder estas preguntas requiere no solo herramientas estadísticas, sino también un cuidado especial en la forma en que se construye y valida el dataset. Por ello, el énfasis del artículo no se limita a los resultados numéricos, sino que incluye una descripción detallada de las decisiones de ingeniería adoptadas en el *pipeline* ETL.

II. Justificación de la Investigación

La motivación central de este trabajo puede resumirse en una idea sencilla: en Arequipa existe suficiente información pública para analizar la relación minería–salud con rigor, pero dicha información aún no se ha articulado en una herramienta analítica integrada y accesible. Lo que sigue es, en esencia, una justificación de por qué valía la pena invertir esfuerzo en construir esa herramienta y en documentar el proceso.

A. Justificación Social y Territorial

Arequipa combina una fuerte dependencia económica de la minería con una historia de conflictos socioambientales vinculados a la percepción de riesgo sanitario. En la práctica, esto se traduce en una dualidad cotidiana: por un lado, la población reconoce los beneficios del canon, el empleo y la inversión; por otro, expresa preocupación por el polvo en suspensión, el ruido, el tránsito pesado y la posible contaminación del aire y del agua.

En este contexto, la discusión pública suele apoyarse en testimonios o casos aislados ("mi vecino se enfermó después de que se expandió la mina"), pero rara vez en series de tiempo comparables o análisis estadísticos reproducibles. La ausencia de evidencia integrada alimenta la polarización: cada actor selecciona los datos que refuerzan su posición, sin un marco común de referencia.

El presente estudio busca reducir esa brecha ofreciendo un análisis basado exclusivamente en datos oficiales del propio Estado peruano. La elección de fuentes públicas no es casual: se trata de minimizar acusaciones de sesgo, tanto a favor como en contra de la actividad minera, y de demostrar que la misma información que ya se reporta puede aprovecharse mejor con herramientas de ingeniería de datos.

B. Justificación Científica

Desde una perspectiva académica, la investigación se justifica porque aborda una pregunta que combina tres dominios tradicionalmente separados: producción minera, epidemiología respiratoria y análisis espacio-temporal de datos. Estudios previos han demostrado la relación entre contaminación por material particulado y enfermedades respiratorias en contextos urbanos o industriales; sin embargo, son menos frecuentes los trabajos que cruzan de manera explícita la producción minera reportada oficialmente con indicadores de salud a nivel distrital.

El proyecto propone un marco metodológico que podría replicarse en otras regiones del país o en otros contextos extractivos, siempre que existan datos abiertos suficientes. De esta forma, el aporte no se limita al caso Arequipa, sino que se extiende como ejemplo de cómo la Ingeniería de Datos puede apoyar la toma de decisiones en salud pública.

C. Propuesta de Valor: de la Percepción a la Evidencia

La propuesta de valor del trabajo puede formularse de forma directa: *pasar de la percepción subjetiva a la evidencia cuantitativa*. Esto no significa negar la validez de la experiencia local, sino complementarla con indicadores que permitan contrastar hipótesis con datos.

El estudio toma una postura deliberadamente neutral: en lugar de asumir que la minería necesariamente enferma o, por el contrario, que no tiene impacto alguno, se plantea una hipótesis nula explícita y se diseña un conjunto de análisis para intentar refutarla. Si la evidencia no alcanza para rechazarla, la interpretación será distinta que si se encontraran patrones claros de asociación.

D. Objetivo Científico y Pregunta Rectora

El objetivo científico principal consiste en evaluar la existencia o ausencia de correlación lineal inmediata entre la producción minera mensual y la incidencia de IRA, acompañando el análisis con la caracterización de posibles efectos diferenciales entre distritos mineros y no mineros.

En términos simples, la pregunta rectora del proyecto puede expresarse así: *"si aumentan las toneladas extraídas en un mes específico, ¿se observa un incremento proporcional de enfermos respiratorios en ese mismo periodo o en los meses siguientes, y dicho efecto es distinto en zonas mineras frente a zonas sin minería?"*.

Responder a esta pregunta con datos reales, aun si la respuesta final resulta ser "no hay correlación inmediata", aporta claridad al debate público y orienta mejor los esfuerzos futuros de monitoreo ambiental y epidemiológico.

III. Justificación de Procesos y Decisiones de Ingeniería

Además de la motivación social y científica, el proyecto se justifica por el valor técnico del *pipeline* desarrollado. En esta sección se explican las decisiones clave de diseño, que no son triviales y condicionan directamente la validez de los resultados.

A. Unificación Temporal: Elección del Mes como Unidad de Análisis

El primer conflicto técnico encontrado fue la diferencia en la granularidad temporal de las fuentes: la GERESA reporta casos de IRA por Semana Epidemiológica (de la 1 a la 52 o 53), mientras que el MINEM publica la producción minera de

forma mensual. Forzar los datos mineros a una escala semanal implicaría introducir supuestos arbitrarios sobre la distribución intramensual de la producción; en cambio, agregar las semanas epidemiológicas a meses calendario es una operación más conservadora desde el punto de vista estadístico.

Por ello, se definió al **mes** como *mínimo común denominador temporal*. En términos operativos, se construyó un mapa semana–mes que asigna cada Semana Epidemiológica al mes en el que se ubica la mayor parte de sus días. De esta manera, la salud se agrega hacia arriba (de semana a mes), mientras que la producción se mantiene en su resolución original.

Esta decisión tiene implicancias directas: el estudio renuncia explícitamente a analizar fenómenos de muy corto plazo (por ejemplo, brotes de pocos días) para ganar consistencia en la comparación minería–salud. El objetivo no es modelar cada brote puntual, sino identificar patrones mes a mes.

B. Normalización Poblacional: de Números Absolutos a Tasas

Un error frecuente en análisis epidemiológicos exploratorios consiste en comparar números absolutos de casos entre distritos de tamaños de población muy distintos. En Arequipa, distritos como Cerro Colorado o Paucarpata concentran decenas de miles de habitantes, mientras que zonas rurales como Pocsi o San Juan de Tarucani tienen poblaciones mucho menores. En tales condiciones, los casos absolutos de IRA tienden a ser mayores en distritos urbanos simplemente porque hay más personas expuestas, no necesariamente porque el riesgo individual sea más alto.

Para evitar este sesgo, se definió la **Tasa de Incidencia de IRA Normalizada**:

$$\text{Tasa}_{IRA} = \left(\frac{\text{Casos Totales en el Mes}}{\text{Población Proyectada del Distrito}} \right) \times 10\,000 \tag{2}$$

El factor de escala 10 000 se eligió por razones de legibilidad: permite que las tasas adopten valores numéricamente manejables y comparables entre distritos sin recurrir a notación científica. Esta métrica convierte el conteo bruto de enfermos en un indicador de riesgo relativo, que sí puede compararse entre distritos grandes y pequeños.

C. Resolución de "Distritos Huérfanos": Limpieza y Emparejamiento de Entidades

Durante el cruce inicial de las bases de datos se identificó que aproximadamente un 30% de los registros de salud no encontraban su par en la base minera, y viceversa. El problema no era la ausencia de actividad, sino la discrepancia en la manera de nombrar los distritos: la GERESA podía registrar "DISTRITO DE YURA" o "YURA (CS)", mientras que el MINEM empleaba simplemente "YURA"; en otros casos se observaban tildes, abreviaturas o diferencias de mayúsculas.

Ignorar estos casos habría introducido un sesgo grave, en particular porque varios de los distritos afectados corresponden precisamente a zonas de interés minero (Yura, Uchumayo, San Juan de Tarucani). Para resolverlo, se implementó un proceso de *resolución de entidades* basado en:

- Normalización de texto a mayúsculas.
- Eliminación de tildes mediante la librería `unidecode`.
- Limpieza de prefijos y sufijos administrativos (*"DISTRITO DE"*, *"(CS)"*, etc.).
- Verificación manual de casos críticos.

El resultado fue la recuperación de la trazabilidad para prácticamente todos los distritos relevantes. Desde el punto de vista de la calidad de datos, este paso fue determinante: sin él, los análisis habrían sugerido artificialmente que en ciertos distritos mineros "no hay enfermos", cuando en realidad el problema era de nomenclatura.

D. Manejo de Valores Faltantes: NaN frente a Cero

Otra decisión de ingeniería clave fue el tratamiento de valores faltantes en los reportes. Se adoptó la siguiente política:

- **Producción minera faltante** $\rightarrow$ se interpreta como producción nula (cero), asumiendo que la ausencia de reporte en un periodo específico implica que no hubo actividad registrada o que la escala del proyecto no afecta de manera significativa el volumen regional.
- **Reporte de salud faltante** $\rightarrow$ se mantiene como `NaN` (no disponible), evitando reemplazarlo arbitrariamente por cero.

Esta distinción obedece a un principio de prudencia epidemiológica: registrar "cero enfermos" cuando simplemente no se recibió información de un centro de salud podría reducir artificialmente la tasa promedio de morbilidad y ocultar posibles problemas de subreporte. Dejar el valor como `NaN` es estadísticamente más honesto, aun cuando implique que ciertos meses o distritos queden fuera de algunos cálculos agregados.

IV. Marco Teórico

El análisis empírico de la relación entre minería e IRA requiere un marco conceptual que articule tres dominios: la epidemiología de las infecciones respiratorias, la dinámica de la contaminación atmosférica en contextos mineros y los principios básicos del análisis espacio-temporal de datos de salud. En esta sección se presenta una síntesis de dichos elementos, con énfasis en aquellos aspectos que resultan directamente relevantes para la región Arequipa.

A. Infecciones Respiratorias Agudas y Vulnerabilidad en Zonas Altoandinas

Las Infecciones Respiratorias Agudas constituyen uno de los principales motivos de consulta y hospitalización en el Perú, particularmente en población infantil y adulta mayor. La literatura epidemiológica resalta que la incidencia de IRA está fuertemente modulada por factores ambientales y sociales: temperatura, humedad, hacinamiento, calidad del aire en interiores y acceso a servicios de salud, entre otros.

En zonas altoandinas como Arequipa, el patrón estacional de las IRA presenta un incremento marcado durante los meses fríos. La combinación de bajas temperaturas nocturnas, viviendas con aislamiento térmico limitado y exposición a

combustión de biomasa en interiores incrementa la susceptibilidad a infecciones de vías respiratorias superiores e inferiores. Este patrón estacional ha sido documentado en reportes del Ministerio de Salud y constituye un punto de partida fundamental: cualquier análisis sobre el rol potencial de la minería debe reconocer que, incluso en ausencia de actividad extractiva, el clima por sí solo genera picos de morbilidad.

B. Contaminación por Material Particulado y Salud Respiratoria

Diversos estudios internacionales han demostrado la asociación entre concentraciones elevadas de material particulado respirable (PM_{10} y $PM_{2.5}$) y mayores tasas de enfermedades respiratorias, exacerbaciones de asma, hospitalizaciones y mortalidad prematura. El mecanismo subyacente incluye la irritación de las vías respiratorias, la inflamación crónica y, en el caso de partículas que contienen metales pesados, la generación de estrés oxidativo.

En contextos mineros, las fuentes de material particulado incluyen la extracción y chancado de roca, el transporte de mineral en camiones de alto tonelaje, las voladuras y, en algunos casos, los procesos de fundición. La dispersión de partículas depende, además, de la dirección e intensidad del viento, de la precipitación y de la topografía local. La hipótesis de trabajo más razonable, por tanto, no es que la minería genere un pico súbito de casos de IRA en un mes específico, sino que contribuya a un aumento del *riesgo basal* de la población expuesta, haciendo que los episodios virales estacionales tengan un impacto más severo.

C. Estacionalidad, Tendencias y Desfasajes

El análisis de series de tiempo en epidemiología distingue entre tres componentes principales: tendencia, estacionalidad y ruido. La tendencia captura cambios de largo plazo; la estacionalidad refleja patrones recurrentes en periodos específicos del año; el ruido agrupa variaciones no explicadas. En la región Arequipa, los datos de IRA muestran una estacionalidad clara ligada al invierno, lo cual introduce un reto metodológico: si no se controla esta estacionalidad, cualquier comparación entre distritos puede confundirse por el simple hecho de que comparten el mismo clima regional.

Además, es posible que ciertos factores ambientales, como la contaminación atmosférica, no se manifiesten de forma instantánea sino con cierto desfasaje temporal. Por ello, la relación entre producción minera y salud podría no ser estrictamente simultánea (mismo mes), sino implicar efectos acumulativos o retardados. En la práctica, la disponibilidad de datos mensuales limita la capacidad de detectar desfasajes finos, pero el marco teórico invita a considerar esta posibilidad al interpretar los resultados.

D. Análisis Espacio-Temporal y Comparación entre Distritos

El componente espacial del análisis es igualmente importante. No basta con evaluar una correlación global entre producción regional e IRA total: es necesario distinguir entre distritos con presencia directa de unidades mineras y distritos sin actividad extractiva, y comparar sus trayectorias epidemiológicas a lo largo del tiempo. Este enfoque permite evaluar si, más allá de la estacionalidad común, existe una diferencia sistemática en el nivel basal de morbilidad entre ambos grupos.

La literatura sobre epidemiología espacial emplea con frecuencia técnicas como la autocorrelación de Moran, la regresión ponderada geográficamente (GWR) o los modelos autorregresivos espaciales (SAR/SEM) para capturar patrones de dependencia entre unidades territoriales. Si bien estos métodos exceden el alcance de la presente fase del proyecto, constituyen la base teórica para las líneas de trabajo futuro que se propondrán en secciones posteriores.

V. Metodología de Datos y Diseño del Análisis Estadístico

Una vez establecido el marco conceptual, el siguiente paso consistió en traducir las preguntas de investigación en un diseño de análisis concreto. Esta sección detalla la construcción del dataset integrado, la definición de variables, los niveles de agregación considerados y los métodos estadísticos utilizados para contrastar la hipótesis nula y explorar patrones espacio–temporales.

A. Construcción del Dataset Integrado

El *pipeline* ETL desarrollado en Python permitió obtener, para cada combinación de *distrito–mes–año*, la siguiente información mínima:

- Producción minera mensual total (en TMF), agregada por distrito.
- Casos de IRA, neumonía y defunciones respiratorias, sumados por mes.
- Población proyectada del distrito para el año correspondiente.
- Etiqueta binaria que indica si el distrito presenta o no actividad minera relevante.

En el caso de las provincias, la información se obtuvo agregando por suma (producción y casos de salud) o por suma ponderada (tasa de incidencia) todos los distritos contenidos en cada provincia.

B. Variables de Estudio

En la Tabla I se resume el conjunto de variables analizadas. Nótese que el énfasis recae en la tasa de incidencia normalizada, más que en los conteos absolutos.

TABLE I
PRINCIPALES VARIABLES EMPLEADAS EN EL ANÁLISIS

Variable	Descripción
Producción_TMF	Producción minera mensual total del distrito (TMF), suma de metales y no metales reportados por MINEM.
Casos_IRA	Total mensual de casos de IRA reportados por GERESA (incluye IRA sin neumonía, neumonías y defunciones asociadas).
Población	Población distrital proyectada por INEI para el año correspondiente.
Tasa_IRA	Casos_IRA normalizados por población y escalados por 10 000 habitantes.
Mes	Identificador de mes (1–12), utilizado para capturar estacionalidad.
Distrito_Minero	Indicador binario (1 = distrito con actividad minera reportada; 0 = distrito sin actividad minera).

C. Niveles de Análisis

Se definieron tres niveles complementarios:

1) **Nivel global**: toda la región Arequipa considerada como un solo sistema; se evalúa la correlación entre producción total e incidencia total.
2) **Nivel provincial**: comparación entre provincias, identificando patrones diferenciados según su grado de actividad minera.
3) **Nivel distrital**: análisis detallado de distritos seleccionados (Yura, Uchumayo, Callalli, San Juan de Tarucani, entre otros), con énfasis en la dinámica local de producción y enfermedad.

Cada nivel aporta una perspectiva distinta: el análisis global permite evaluar la fuerza de la señal promedio; el provincial revela heterogeneidades intermedias; el distrital se acerca a la escala en la que se experimentan los impactos ambientales y sanitarios.

D. Procedimientos Estadísticos

Para contrastar la hipótesis nula y caracterizar la relación minería–IRA se aplicaron los siguientes procedimientos:

1) Correlación de Pearson: Se utilizó el coeficiente de Pearson para evaluar la existencia de una relación lineal entre producción mensual (en escala logarítmica) y tasa de IRA. La transformación logarítmica de la producción se empleó para reducir la asimetría en la distribución de TMF, típica en datos económicos de magnitud muy variable.

2) Correlación de Spearman: Dado que no se puede garantizar la normalidad de las variables, se calculó adicionalmente la correlación de Spearman, que se basa en rangos y permite explorar relaciones monótonas no necesariamente lineales. La compatibilidad de resultados entre ambos coeficientes refuerza la interpretación.

3) Comparación de Medias entre Distritos Mineros y No Mineros: Para evaluar si los distritos con actividad minera presentan una incidencia basal de IRA distinta a los distritos sin actividad minera se aplicó una prueba de diferencia de medias. Dado que los tamaños de muestra y las varianzas pueden ser diferentes entre grupos, se adoptó la versión robusta de la prueba *t* de Student con corrección de Welch. El contraste se realizó tanto sobre la tasa mensual promedio como sobre indicadores agregados por año.

4) Análisis de Estacionalidad y Desfasajes: La estacionalidad se exploró visualmente mediante series temporales y gráficas de tendencia por mes. Además, se evaluaron posibles desfasajes entre producción y tasa de IRA mediante el análisis de correlación cruzada (*cross-correlation*) entre la serie de producción y la serie de tasas, considerando retardos de uno a tres meses. Si bien la resolución mensual limita el poder de detección de efectos rápidos, este análisis permite identificar patrones diferidos de primer orden.

5) Construcción del Indicador de "Offset": Para formalizar la brecha observada entre distritos mineros y no mineros, se construyó un indicador de *offset* definido como la diferencia promedio mensual entre las tasas de ambos grupos, manteniendo fija la estacionalidad. Este indicador se interpreta como una aproximación al *riesgo basal adicional* asociado a residir en un distrito minero, independientemente del patrón climático.

VI. RESULTADOS

En esta sección se presentan los principales hallazgos del análisis cuantitativo. Se sigue la estructura de niveles definida anteriormente: primero se discute la relación global, luego los patrones provinciales y, finalmente, los comportamientos distritales y el análisis comparativo mineros–no mineros.

A. Relación Global Producción–IRA

La Figura 1 muestra el diagrama de dispersión entre la producción minera total mensual (en escala logarítmica) y la tasa de IRA regional correspondiente al mismo mes.

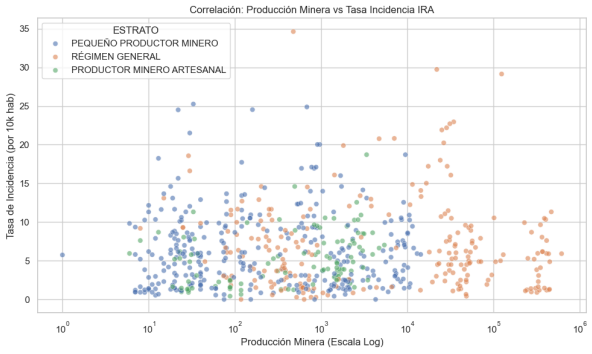

Fig. 1. Dispersión entre producción minera mensual (log-TMF) y tasa de incidencia de IRA en Arequipa.

La nube de puntos es claramente difusa, sin una tendencia ascendente o descendente evidente. Los coeficientes de correlación de Pearson y Spearman calculados sobre este conjunto de datos se encuentran muy próximos a cero, lo que indica ausencia de relación lineal o monótona fuerte entre ambas variables. En términos de la hipótesis nula planteada en la Introducción, este resultado sugiere que, al menos a nivel regional y en escala mensual, no se observa un efecto inmediato de la producción minera sobre la incidencia de IRA.

B. Comportamiento Provincial

Para explorar posibles heterogeneidades intermedias se construyeron gráficos de tendencia a nivel provincial. La Figura 2 presenta la evolución de la tasa de IRA promedio por provincia, mientras que las Figuras 3–7 muestran, para cada provincia seleccionada, la serie conjunta de producción minera y tasa de IRA.

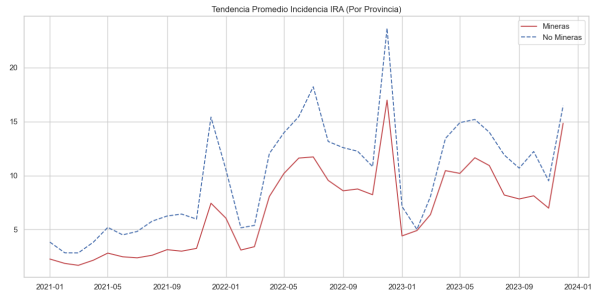

Fig. 2. Tendencia temporal de la tasa de IRA a nivel provincial (Arequipa, Camaná, Caravelí, Caylloma e Islay).

En todas las provincias se observa un patrón estacional muy similar: las tasas aumentan durante los meses de otoño e invierno y disminuyen en primavera–verano. Esta sincronía provincial refuerza la hipótesis de que el clima es el principal modulador de los picos de IRA, por encima de factores locales específicos.

Las Figuras 3 a 7 superponen la producción minera provincial (barras o línea) y la tasa de IRA (línea), permitiendo visualizar la relación local entre ambas series.

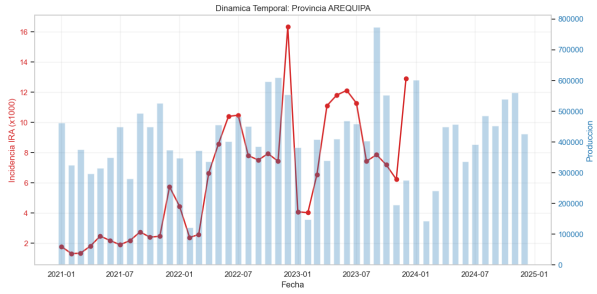

Fig. 3. Provincia de Arequipa: producción minera mensual vs. tasa de IRA.

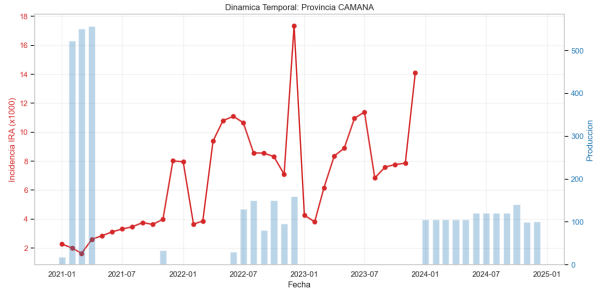

Fig. 4. Provincia de Camaná: producción minera mensual vs. tasa de IRA.

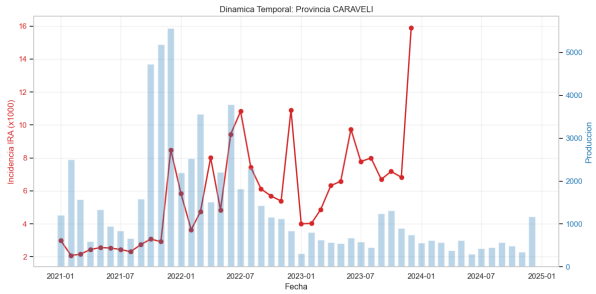

Fig. 5. Provincia de Caravelí: producción minera mensual vs. tasa de IRA.

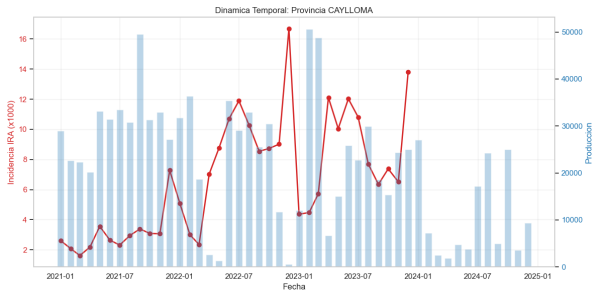

Fig. 6. Provincia de Caylloma: producción minera mensual vs. tasa de IRA.

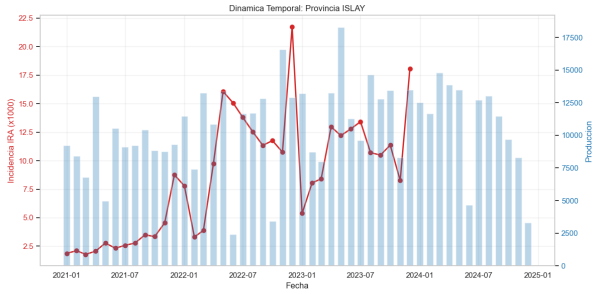

Fig. 7. Provincia de Islay: producción minera mensual vs. tasa de IRA.

En ninguno de los casos se aprecia una relación directa consistente entre picos de producción y picos de enfermedad. En cambio, lo que se observa es nuevamente la predominancia del ciclo estacional en la curva epidemiológica, relativamente independiente de las fluctuaciones de producción.

C. Análisis Distrital: Yura, Uchumayo, Callalli y San Juan de Tarucani

El análisis distrital profundiza en unidades territoriales directamente expuestas a operaciones mineras. Las Figuras 8 a 11 muestran los gráficos de doble eje para cuatro distritos emblemáticos.

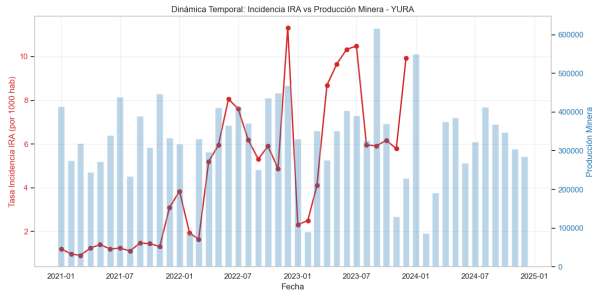

Fig. 8. Distrito de Yura: producción minera mensual vs. tasa de IRA.

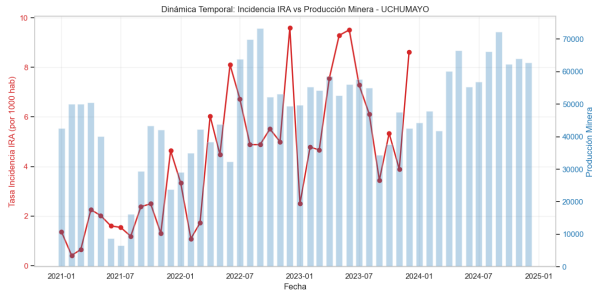

Fig. 9. Distrito de Uchumayo: producción minera mensual vs. tasa de IRA.

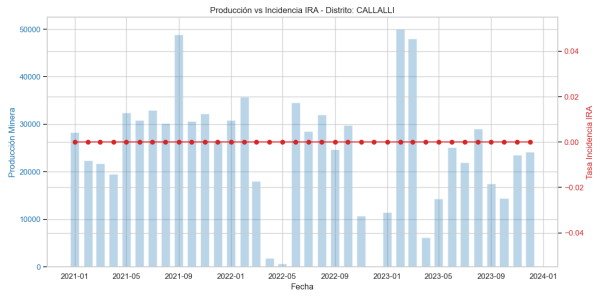

Fig. 10. Distrito de Callalli: producción minera mensual vs. tasa de IRA.

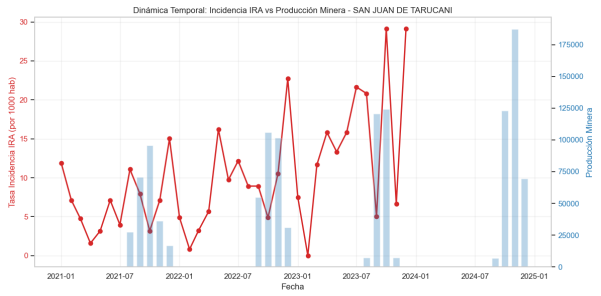

Fig. 11. Distrito de San Juan de Tarucani: producción minera mensual vs. tasa de IRA.

En los cuatro distritos se observa el mismo fenómeno: la curva de tasa de IRA sigue un ciclo anual similar al del resto de la región, mientras que la producción minera presenta picos y valles que no coinciden de manera sistemática con los picos de enfermedad. Este patrón de *asincronía* respalda la conclusión de que la actividad extractiva mensual no explica por sí sola los brotes agudos de IRA.

D. Comparación Mineros vs No Mineros y Fenómeno de "Offset"

Para evaluar diferencias estructurales entre distritos mineros y no mineros, se construyeron dos cohortes: una que agrupa a los distritos con actividad minera y otra que reúne a los distritos sin dicha actividad. La Figura 12 muestra un diagrama de caja comparando la distribución de tasas de IRA entre ambos grupos, mientras que la Figura 13 presenta la evolución temporal de sus promedios.

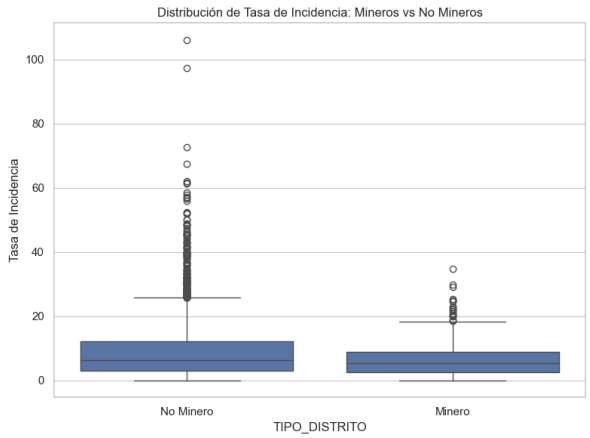

Fig. 12. Distribución comparativa de la tasa de IRA en distritos mineros vs. no mineros.

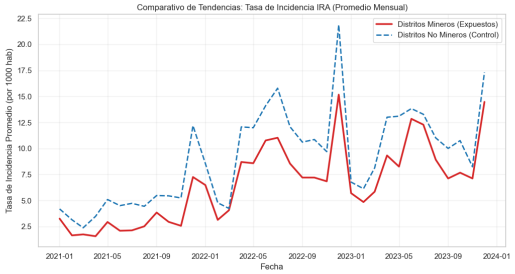

Fig. 13. Tendencias temporales de la tasa de IRA en distritos mineros y no mineros.

El diagrama de caja sugiere que la mediana y gran parte de la distribución de las tasas en distritos mineros se sitúa por encima de la de los distritos no mineros. La curva de tendencias muestra que ambas series comparten la misma estacionalidad (suben y bajan al mismo tiempo), pero la serie de distritos mineros se mantiene de manera consistente algunos puntos por encima: este es el *offset* o brecha basal identificada.

La prueba de diferencia de medias, aplicada sobre las tasas mensuales promediadas por grupo, indica que esta diferencia es estadísticamente significativa bajo un nivel de confianza razonable. Aunque la magnitud de la brecha no es suficiente

para atribuir causalidad directa a la minería, sí sugiere que la población de distritos mineros enfrenta un riesgo respiratorio ligeramente mayor durante todo el año.

VII. Discusión

Los resultados permiten responder de manera matizada la pregunta que motivó el proyecto. En primer lugar, los análisis de correlación global y los gráficos de dispersión muestran que no existe una asociación lineal fuerte entre la producción minera mensual y la tasa de IRA en la región Arequipa. En términos de la hipótesis nula planteada, la evidencia empírica respalda la conclusión de que los picos mensuales de producción no generan, por sí solos, picos simultáneos de enfermedad respiratoria.

En segundo lugar, el análisis provincial y distrital confirma que el componente dominante en la dinámica de las IRA es la estacionalidad climática. Tanto los distritos mineros como los no mineros experimentan brotes más intensos durante los meses fríos, lo que coincide con la literatura epidemiológica sobre zonas altoandinas. Este hallazgo es importante porque contextualiza la percepción social: la coincidencia entre invierno y ciertas actividades mineras puede llevar a interpretar como causalidad lo que, en gran medida, es un efecto del clima.

Sin embargo, la investigación no se limita a concluir que "la minería no influye"; de hecho, los resultados descubren un patrón más sutil. La comparación entre distritos mineros y no mineros revela una diferencia sistemática en el nivel basal de la tasa de IRA: incluso cuando ambas series comparten el mismo ciclo estacional, la curva de distritos mineros se encuentra consistentemente por encima. Esta brecha, aunque moderada, es robusta y estadísticamente significativa.

Desde una perspectiva epidemiológica, este *offset* puede interpretarse como un aumento de vulnerabilidad crónica en la población residente en zonas mineras. Entre los posibles mecanismos que podrían explicar esta situación se encuentran:

- Una exposición sostenida a material particulado en el ambiente, producto de la actividad extractiva y del tránsito de vehículos pesados.
- Condiciones socioeconómicas específicas de las zonas mineras (vivienda precaria, acceso desigual a servicios de salud, mayores niveles de estrés laboral).
- Interacciones entre polvo ambiental y factores de riesgo ya presentes (humo de leña, condiciones climáticas extremas).

El estudio no cuenta con datos suficientes para discriminar entre estas explicaciones, pero sí establece un punto de partida sólido: la minería no parece ser el motor inmediato de los picos estacionales, pero podría estar relacionada con una elevación sostenida del riesgo respiratorio.

Finalmente, es importante resaltar el valor de la metodología. Más allá de los resultados específicos para Arequipa, el *pipeline* de integración de datos demuestra que es posible auditar la relación entre minería y salud utilizando únicamente datos abiertos y herramientas de software libre. Esto abre la puerta a replicar el enfoque en otras regiones del país o en otros sectores productivos.

VIII. Limitaciones del Estudio

Como todo estudio observacional basado en datos secundarios, el presente trabajo presenta varias limitaciones que deben reconocerse explícitamente:

- **Resolución temporal**: la elección del mes como unidad de análisis, aunque metodológicamente coherente, impide detectar efectos de corto plazo (por ejemplo, incrementos de partículas en días específicos).
- **Calidad de los datos de salud**: la tasa de IRA depende de la oportunidad y exhaustividad del reporte de los establecimientos de salud. Es posible que exista subregistro, especialmente en zonas rurales de difícil acceso.
- **Ausencia de datos ambientales directos**: no se contó con series temporales completas de PM_{10}, $PM_{2.5}$, temperatura y humedad para los distritos analizados. Esto limita la capacidad de modelar explícitamente el rol de la contaminación atmosférica.
- **Causalidad vs. correlación**: el enfoque del estudio es correlacional; incluso si se observaran asociaciones fuertes, no sería posible atribuir causalidad sin un diseño experimental o cuasi-experimental más complejo.
- **Heterogeneidad interna de la actividad minera**: el análisis considera la producción total por distrito, sin diferenciar entre tipos de minerales, tecnologías empleadas o prácticas de mitigación ambiental, que podrían tener impactos muy distintos.

Estas limitaciones no invalidan los resultados, pero sí delimitan su alcance y señalan la necesidad de complementar este tipo de análisis con información adicional.

IX. Trabajo Futuro

A partir de la experiencia adquirida en este proyecto, se identifican varias líneas de trabajo futuro:

- **Integración de datos ambientales**: incorporar series históricas de concentración de PM_{10} y $PM_{2.5}$, así como variables meteorológicas (temperatura, humedad, velocidad y dirección del viento), permitiría construir modelos multivariados más robustos.
- **Análisis espacial avanzado**: aplicar estadísticas de autocorrelación espacial (como el índice de Moran) y modelos de regresión espacial (SAR, SEM, GWR) para capturar la dependencia entre distritos vecinos y evaluar la propagación geográfica del riesgo.
- **Desarrollo de una plataforma interactiva**: empaquetar el *pipeline* ETL y los resultados en un tablero interactivo (*dashboard*) accesible para autoridades y ciudadanía, favoreciendo la transparencia y la participación informada.
- **Modelos predictivos**: explorar técnicas de aprendizaje automático (por ejemplo, bosques aleatorios o redes recurrentes) para predecir la carga de IRA a partir de variables climáticas, productivas y ambientales, evaluando distintos escenarios de intervención.

- **Vinculación con estudios clínicos y comunitarios**: complementar el análisis cuantitativo con encuestas, mediciones clínicas y monitoreo de calidad de aire a nivel domiciliario en distritos seleccionados, con el fin de validar hipótesis sobre mecanismos de exposición y vulnerabilidad.

En conjunto, estas líneas de trabajo permitirían avanzar desde el diagnóstico cuantitativo inicial hacia una comprensión más profunda y accionable de la relación entre minería y salud en Arequipa.

X. Conclusiones

El presente estudio desarrolló y aplicó un *pipeline* de Ingeniería de Datos para integrar información minera, epidemiológica y demográfica de la región Arequipa durante el periodo 2021–2023. A partir de este dataset unificado, se evaluó la posible relación entre producción minera mensual e incidencia de Infecciones Respiratorias Agudas, considerando distintos niveles de análisis (regional, provincial y distrital).

Los principales hallazgos pueden resumirse en los siguientes puntos:

1) A nivel regional y provincial, los coeficientes de correlación entre producción minera mensual y tasa de IRA son cercanos a cero, lo que respalda la hipótesis nula de ausencia de relación lineal inmediata entre ambas variables.
2) La estacionalidad climática emerge como el factor dominante en la dinámica de las IRA: todos los distritos, mineros y no mineros, presentan picos de morbilidad durante los meses fríos, independientemente de las fluctuaciones de producción.
3) Aun cuando los picos estacionales son compartidos, los distritos con actividad minera muestran, en promedio, una tasa de IRA ligeramente superior a la de los distritos sin minería, configurando un *offset* o brecha basal que sugiere una posible vulnerabilidad crónica asociada al contexto minero.
4) El *pipeline* ETL desarrollado demuestra que es posible auditar la relación minería–salud utilizando exclusivamente datos abiertos y herramientas de software libre, lo que representa un aporte metodológico replicable en otras regiones y contextos.

En conjunto, los resultados invitan a abandonar explicaciones simplistas. La minería no parece ser el desencadenante directo de los brotes agudos de IRA, pero sí podría estar contribuyendo a un aumento sostenido del riesgo respiratorio en las poblaciones expuestas. Confirmar o refinar esta hipótesis requerirá incorporar datos ambientales y diseños de investigación más complejos, pero el presente trabajo establece una base cuantitativa sobre la cual dichos esfuerzos pueden construirse.

Agradecimientos

Los autores agradecen al equipo docente del curso de la Escuela Profesional de Ingeniería de Sistemas de la UNSA por la orientación metodológica durante el desarrollo del proyecto, así como a las instituciones productoras de datos abiertos (MINEM, GERESA, INEI) cuyo esfuerzo de transparencia hizo posible este estudio.

References

[1] Ministerio de Energía y Minas del Perú (MINEM), "Boletín Estadístico Minero 2021–2023," Lima, Perú.
[2] Gerencia Regional de Salud de Arequipa (GERESA), "Reportes de la Sala Situacional de Infecciones Respiratorias Agudas," 2021–2023.
[3] Instituto Nacional de Estadística e Informática (INEI), "Proyecciones de Población Distrital 2018–2025," Lima, Perú.
[4] Organización Mundial de la Salud (OMS), "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," Informe Técnico, 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, "Repositorio del Proyecto HADS: Análisis Minería-Salud en Arequipa," GitHub, 2025. [En línea]. Disponible: https://github.com/ycozco/HADS_Bumblebees.