Exploración algorítmica de relaciones entre calidad de aire y bienestar

Satu Elisa Schaeffer
Facultad de Ingeniería Mecánica y Eléctrica

PAICYT 2020

Resumen

Buscamos desarrollar una plataforma interactiva con técnicas de ciencia de datos a base de mediciones de calidad de aire que incluyen información climática para automáticamente explorar relaciones multifactoriales entre indicadores de calidad de aire y datos de bienestar poblacional como por ejemplo consultas en centros de salud, combinando información ambiental y socioeconómico en lo temporal y lo espacial. La forma tradicional de buscar por relaciones de este tipo es uno por uno, relacionando una enfermedad específica con un contaminante particular en un marco de tiempo establecido.

Nosotros queremos aplicar la minería de datos para identificar relaciones entre múltiples atributos sobre varios plazos temporales, ya que por ejemplo cáncer pulmonar puede tardar más en presentarse que asma en zonas de aire contaminado y depresión se puede relacionar a una combinación de condiciones climáticas y de contaminación a corto y largo plazo.

1. Introducción

Actualmente contamos con más de una década de datos detallados de calidad de aire, capturados una vez por hora en múltiples puntos del área metropolitana por el (*Sistema Integral de Monitoreo Ambiental*, SIMA)[1], cuya directora **Armandina Valdez Cavazos**, con integrantes de su equipo como por ejemplo Gerardo Argullín García y Jair Rafael Carrillo Avila, colabora con nuestro equipo de trabajo. Además, a través de SIMA, contamos con canales comunicación a otras entidades dentro del gobierno estatal.

La meta de la investigación es la creación de una herramienta que permita que profesionales de la salud y tomadores de decisión del gobierno, junto con miembros del público en general, puedan fácilmente analizar y visualizar relaciones entre la calidad de aire y datos diversos de bienestar poblacional según su preferencia de tal forma que puedan contar con información clara del significado estadístico de las relaciones presentes y del plazo de tiempo que requieren en manifestarse. Los resultados se reportarán en publicaciones internacionales en revistas indizadas y todo el software creado se hará disponible con una licencia de código abierto para uso libre sin fines de lucro.

[1] `http://aire.nl.gob.mx/`

El conjunto de técnicas teóricos para el proyecto incluye la teoría de grafos, ecuaciones diferenciales y la estadística inferencial, mientras lo algorítmico involucra técnicas de aprendizaje máquina incluyendo clasificación y pronóstico. La interfaz de usuario será un sitio web interactivo con gráficas animadas y la opción de descargar datos en una hoja de cálculo para análisis adicional posterior con las herramientas de su selección.

La presente propuesta sigue la investigación iniciada en el proyecto PAICYT IT512-15 *Herramientas computacionales para análisis epidemiológico multifactorial* y es asociada a la propuesta CF-MI-20191001115100876-2035 que se encuentra en evaluación en la convocatoria 2019 de **Ciencia de Frontera** de CONACYT.

2. Antecedentes

Problemas relacionadas a la calidad de aire son complejos y afectan a la sociedad en múltiples formas, especialmente a largo plazo. Numerosos estudios publicados establecen enlaces entre contaminantes y riesgos de salud — crónicos y agudos — (Allen y cols., 2016), en adición a la literatura tradicional que establece lazos entre condiciones climáticas y la salud poblacional (Costilla Esquivel y cols., 2014).

Nuestra ciudad recibe frecuentemente mala prensa por su calidad de aire. Trabajos científicos que puedan aclarar causas y consecuencias de la contaminación son vitales para facilitar la planeación y la toma de decisión para atender a esta crisis; el diálogo público pasivo del estilo "alguien debe tomar acción" impide que se tome acción concreta mientras nadie sabe con certeza quién debe hacer ni qué ni en dónde ya que no se ha modelado a detalle ni de dónde provienen las contaminantes ni qué efectos tienen. Además se observan cambios en los patrones de calidad de aire debidos al cambio climático por su efecto en los fenómenos atmosféricos.

Proyectos como esta propuesta pueden apoyar y fortalecer las prácticas de recolección, procesamiento, difusión y análisis de datos de calidad aire con un soporte adecuado tecnológico y científico para que se puedan llegar a conclusiones reproducibles y así a una toma de decisiones sana y sólida a través de leyes, regulaciones, normas y otras políticas públicas con una comunicación abierta y trasparente al público en general sobre el tema.

3. Objetivos y metas

El **objetivo general** es identificar relaciones multifactoriales entre mediciones de calidad de aire y conjuntos de datos de bienestar poblacional, en específico de salud pública, para la zona metropolitana de Monterrey, Nuevo León, consolidando una metodología reproducible y automatizada que permite generalizar el estudio posteriormente a otras zonas y otros tipos de datos de bienestar.

Los *objetivos específicos* para el presente año son

Modelado matemático Diseñar, implementar, y validar un modelo matemático para series de tiempo georeferenciados tipo multiatributo y multiubicación tridimensional.

Datos abiertos Apoyar la captura, el preprocesamiento y el análisis de datos de calidad de aire para la zona metropolitana de Monterrey, en colaboración con SIMA.

Análisis automatizado Implementar un prototipo de una plataforma de ciencia de datos con herramientas de aprendizaje máquina para explorar relaciones multifactoriales entre calidad de aire y datos de bienestar poblacional.

Visualización científica Implementar un prototipo de un servicio web que visualiza con animaciones interactivas e una indicación clara de significancia estadística dichas relaciones.

Mejora de políticas pública Aportar datos y conclusiones al gobierno estatal para apoyar su proceso de planeación y toma de decisiones relacionados a la calidad de aire.

3.1. Hipótesis

Al combinar series de tiempo georeferenciados de datos multiatributo de calidad de aire con datos de consultas en centros de salud públicos permite la exploración automatizada de relaciones multifactoriales entre la calidad de aire y el bienestar en términos de salud pública.

4. Metodología

En colaboración previa con SIMA, hemos preprocesado, anotado y analizado promedios por hora de 16 variables capturadas en 13 estaciones de calidad de aire, incluyendo PM 2.5, PM 10, ozono y diversos gases así como condiciones climáticas: radiación solar, temperatura, humedad, precipitación, velocidad y dirección de viento. Conocemos la ubicación precisa de cada estación de monitoreo.

Una vez que se cuenta con el modelo interpolado de calidad de aire, se incorporan los datos de salud pública, usando las categorías de CIE de los diagnósticos junto con la ubicación del centro de salud en el cual se realizó la consulta para establecer presencia o ausencia de enfermedades de manera georeferenciada como series de tiempo por códigos CIE, así como se hizo en la tesis de maestría en J. A. Benavides Vázquez (2019) pero ahora combinándolo con los datos de calidad de aire y la georeferenciación.

En el proyecto propuesto, combinamos, adaptamos y extendemos métodos propuestos en la literatura para interpolar datos georeferenciados (Appice, Ciampi, Fumarola, y Malerba, 2014; Dobesch, Dumolard, y Dyras, 2013; Kobler y cols., 2007), para modelar series de tiempo con ecuaciones diferenciales (Chen, Yang, Meng, Zhao, y Abraham, 2011; Eisenhammer, Hübler, Packard, y Kelso, 1991; Xue y Lai, 2018) como se ha hecho anteriormente con datos epidemiológicos de un sólo factor (Lega y Brown, 2016; Miao, Wang, Zhang, Wang, y Pradeep, 2017; Ponciano y Capistrán, 2011; Rasmussen, Ratmann, y Koelle, 2011).

La meta es diseñar e implementar un método que permita la interpolación de series de tiempo multifactoriales de datos georeferenciados de calidad de aire dentro de un equivalente tridimensional del polígono convexo envolviendo los puntos de medición (latitud, longitud y altitud sobre el nivel de mar). Queremos tomar en cuenta la forma geográfica de la zona (extraída de Google Earth con una rejilla de coordenadas) para poder determinar a un momento

de tiempo arbitrario entre la primera y la última medición el nivel de presencia instantánea o acumulada de tipos específicos de contaminantes en el aire.

Luego, con referencias cruzadas a datos de bienestar poblacional, como por ejemplo los datos de salud propuestos para este primer estudio, podramos aplicar métodos de inteligencia artificial para construir modelos multifactoriales que permiten determinar cuáles atributos de bienestar cuentan con relaciones estadísticamente significativas con distintas variables de calidad de aire para distintos ventanas de tiempo y retrasos.

5. Infraestructura y apoyo técnico disponible

El cubículo del investigador responsable en el CIDET cuenta con una iMac adquirida por el proyecto PAICYT 2015 que es suficiente para el trabajo por realizarse durante el 2020. Se ha solicitado equipo adicional en la propuesta a CONACYT que sigue en evaluación en este momento para los trabajos computacionales de fases posteriores en 2021 y 2022.

6. Participantes

Además del tesista de doctorado *Alberto Benavides* y el tesista de licenciatura (por seleccionar), se cuenta con la participación de las siguientes tres investigadoras de la FIME; el equipo de trabajo cuenta con resultados previos en el pronóstico y predicción (Rodríguez y Garza Villarreal, 2019; Schaeffer y Rodríguez Sánchez, 2020) y agrupamiento (J. A. Benavides Vázquez, 2019) de series de tiempo igual, el modelado matemáticos de fenómenos de propagación con autómatas celulares (L. A. Benavides Vázquez, Alcalá, Almaguer, Schaeffer, y Berrones Santos, 2018) y grafos (Garza Villarreal y Schaeffer, 2019), análisis multifactorial (Ceballos, Garza Villarreal, y Cantu, 2018) y multiobjetivo (Arellano Arriaga, Molina, Schaeffer, Álvarez Socarrás, y Martínez Salazar, 2019) igual como plataformas tecnológicas para comunicar información al público (Schaeffer y cols., 2018)

Responsable Satu Elisa Schaeffer

- Grado académico: Doctora en Ciencias en Tecnología (Ciencia e Ingeniería de la Computación)
- Nivel en el SNI: 1
- Últimas cinco publicaciones internacionales indizadas:
  1. Schaeffer y Rodríguez Sánchez (2020)
  2. Garza Villarreal y Schaeffer (2019)
  3. Arellano Arriaga y cols. (2019)
  4. L. A. Benavides Vázquez y cols. (2018)
  5. Schaeffer y cols. (2018)

Colaboradora Sara Elena Garza Villarreal

- Grado académico: Doctora en Tecnologías de Información y Comunicaciones (Sistemas Inteligentes)
- Nivel en el SNI: 1
- Últimas cinco publicaciones internacionales indizadas:
  1. Garza Villarreal y Schaeffer (2019)
  2. Rodríguez y Garza Villarreal (2019)
  3. Schaeffer y cols. (2018)
  4. Ceballos y cols. (2018)
  5. Cavazos y Garza Villarreal (2018)

Colaboradora Sara Verónica Rodríguez Sánchez

- Grado académico: Doctora Europea (Ingeniería)
- Nivel en el SNI: 1
- Últimas cinco publicaciones internacionales indizadas:
  1. Schaeffer y Rodríguez Sánchez (2020)
  2. Escalante y cols. (2019)
  3. Rodríguez Sánchez, Pla-Aragones, y De Castro (2018)
  4. Zhu, Kumar, Rodríguez Sánchez, y Sriskandarajah (2015)
  5. Rodríguez Sánchez, Plà, y Faulin (2014)

Contamos con el apoyo del SIMA y la colaboración del Dr. **José Gerardo Velasco Castañón**, investigador médico jubilado, presidente de una sociedad civil cuyas metas se alinean con las del proyecto propuesto. Además tenemos un contacto en el Hospital Universitario con investigadores quienes han trabajado con la contaminación de aire.

7. Formación de recursos humanos

M.C. *José Alberto Benavides Vázquez* inició su trabajo doctoral en el tema en enero 2020. Adicionalmente se espera involucrar al proyecto propuesto un tesista de licenciatura a partir de agosto 2020, quien defendería su tesis a inicios del 2021. Por el momento no se puede prever que alguien defienda una tesis en este tema antes de que *concluya* el presente año, sin embargo, se avanzarán trabajos de tesis durante ese tiempo.

8. Calendarización de actividades

8.1. Primera administración

Con la participación de los alumnos de verano científico, el periodo **mayo–agosto** consiste en exploración inicial de relaciones multifactoriales entre datos de calidad de aire y registros de consultas de centros de salud.

- El estudiante de doctorado se concentra en revisar literatura relacionada e inicia la implementación de los prototipos.
- La investigadora responsable y el estudiante de doctorado diseñan los modelos matemáticos necesarios.
- La investigadoras participantes y el estudiante de doctorado inician la redacción de un artículo tipo revisión (inglés: *survey*) para su futura publicación en una revista indizada internacional.
- El estudiante de licenciatura diseña e implementa su prototipo de visualización.
- Los alumnos de verano científico utilizará Python y R en el análisis de datos; las visualizaciones interactivas de demostración se prepararán con HTML5, CSS3 y JavaScript.

8.2. Segunda administración

En el periodo **septiembre–diciembre** el enfoque fuerte es la *tesis de licenciatura*, cuya comité será formada por el equipo de trabajo, con la responsable como presidente y las colaboradoras como miembros de comité.

- La investigadora responsable y el estudiante de doctorado validan los modelos matemáticos desarrollados.
- El estudiante de doctorado se concentra en finalizar los prototipos y presentar resultados preliminares en un congreso internacional.
- El estudiante de licenciatura redacta la tesis.
- Las investigadoras participantes y el estudiante de doctorado concluyen la redacción del *survey* y la someten en evaluación en una revista indizada internacional.

9. Resultados esperados

Durante el presente año se redacta un artículo tipo *survey* para someter a una revista internacional indizada en coautoría entre los investigadores participantes y el estudiante de doctorado; será tentativamente para la revista de Springer llamada *Environmental and Ecological Statistics*, en acceso abierto si se llegan a recibir fondos de CONACYT (por el alto costo de dicha modalidad).

Los primeros prototipos de software se incluirán en un repositorio público de código abierto. Se documentarán en un artículo en congreso internacional que será presentado por el tesista de doctorado.

Se iniciará una tesis de licenciatura con defensa estimada en febrero 2021 (no será posible iniciar trámites de acta sin kárdex completo, y la unidad de aprendizaje de Investigación se concluye a inicios de diciembre del 2020). La tesis doctoral de Alberto Benavides se inicia en el 2020 y se estima defenderla a inicios del 2023.

Referencias

Allen, R. T., Hales, N. M., Baccarelli, A., Jerrett, M., Ezzati, M., Dockery, D. W., y III, C. A. P. (2016). Countervailing effects of income, air pollution, smoking, and obesity on aging and life expectancy: population-based study of U.S. counties. *Environmental Health*, *15* (1), 86. doi: doi:10.1186/s12940-016-0168-2

Appice, A., Ciampi, A., Fumarola, F., y Malerba, D. (2014). *Data mining techniques in sensor networks: Summarization, interpolation and surveillance*. Springer. doi: doi:10.1007/978-1-4471-5454-9

Arellano Arriaga, N. A., Molina, J., Schaeffer, S. E., Álvarez Socarrás, A. M., y Martínez Salazar, I. A. (2019, junio). A bi-objective study of the minimum latency problem. *Journal of Heuristics*, *25* (3), 431–454. doi: doi:10.1007/s10732-019-09405-0

Benavides Vázquez, J. A. (2019). *Agrupamiento no supervisado de series de tiempo epidemiológicas de méxico entre 2005 y 2015* (Tesis de Master, FIME, UANL, Nuevo León, Mexico). Descargado de `https://elisa.dyndns-web.com/students/tesis/msc_benavides2019.pdf`

Benavides Vázquez, L. A., Alcalá, M., Almaguer, F. J., Schaeffer, S. E., y Berrones Santos, J. A. (2018). Cellular automaton model of a wastewater treatment process. *Journal of Cellular Automata*, *13* (5–6), 407–428.

Cavazos, R., y Garza Villarreal, S. E. (2018). Learning models for student performance prediction. En *Advances in computational intelligence* (pp. 171–182). doi: doi:10.1007/978-3-030-02840-4_14

Ceballos, H. G., Garza Villarreal, S. E., y Cantu, F. J. (2018). Factors influencing the formation of intra-institutional formal research groups: group prediction from collaboration, organisational, and topical networks. *Scientometrics*, *114*, 181–216. doi: doi:10.1007/s11192-017-2561-1

Chen, Y., Yang, B., Meng, Q., Zhao, Y., y Abraham, A. (2011, enero). Time-series forecasting using a system of ordinary differential equations. *Information Sciences*, *181*, 106–114. doi: doi:10.1016/j.ins.2010.09.006

Costilla Esquivel, A., Corona-Villavicencio, F., Velasco-Castañón, J. G., Medina de la Garza, C. E., Martínez-Villarreal, R. T., Cortes-Hernández, D. E., ... González Farías, G. (2014). A relationship between acute respiratory illnesses and weather. *Epidemiology and Infection*, *142* (7), 1375–1383. doi: doi:10.1017/S0950268813001854

Dobesch, H., Dumolard, P., y Dyras, I. (Eds.). (2013). *Spatial interpolation for climate data: The use of GIS in climatology and meteorology*. John Wiley & Sons. doi: doi:10.1002/9780470612262

Eisenhammer, T., Hübler, A., Packard, N., y Kelso, J. A. S. (1991). Modeling experimental time

series with ordinary differential equations. *Biological Cybernetics*, *65*, 107-112. doi: doi:10.1007/BF00202385

Escalante, H. J., Rodríguez Sánchez, S. V., Jiménez Lizárraga, M., Morales Reyes, A., de la Calleja, J., y Vazquez, R. (2019). Barley yield and fertilization analysis from uav imagery: a deep learning approach. *International Journal of Remote Sensing*, *40* (7), 2493–2516. doi: doi:10.1080/01431161.2019.1577571

Garza Villarreal, S. E., y Schaeffer, S. E. (2019, noviembre). Community detection with the label propagation algorithm: A survey. *Physica A*, *534*, 122058. doi: doi:10.1016/j.physa.2019.122058

Kobler, A., Pfeifer, N., Ogrinc, P., Todorovski, L., Oštir, K., y Džeroski, S. (2007). Repetitive interpolation: A robust algorithm for DTM generation from aerial laser scanner data in forested terrain. *Remote Sensing of Environment*, *108* (1), 9–23. doi: doi:10.1016/j.rse.2006.10.013

Lega, J., y Brown, H. E. (2016, diciembre). Data-driven outbreak forecasting with a simple nonlinear growth model. *Epidemics*, *17*, 19–26. doi: doi:10.1016/j.epidem.2016.10.002

Miao, A., Wang, X., Zhang, T., Wang, W., y Pradeep, B. S. A. (2017, diciembre). Dynamical analysis of a stochastic SIS epidemic model with nonlinear incidence rate and double epidemic hypothesis. *Advances in Differential Equations*, *2017*, 226. doi: doi:10.1186/s13662-017-1289-9

Ponciano, J. M., y Capistrán, M. A. (2011). First principles modeling of nonlinear incidence rates in seasonal epidemics. *PLOS Computational Biology*, *7* (2), e1001079. doi: doi:10.1371/journal.pcbi.1001079

Rasmussen, D. A., Ratmann, O., y Koelle, K. (2011). Inference for nonlinear epidemiological models using genealogies and time series. *PLOS Computational Biology*, *7* (8), e1002136. doi: doi:10.1371/journal.pcbi.1002136

Rodríguez, F. M., y Garza Villarreal, S. E. (2019). Predicting emotional intensity in social networks. *Journal of Intelligent & Fuzzy Systems*, *36*, 4709–4719. doi: doi:10.3233/JIFS-179020

Rodríguez Sánchez, S. V., Plà, L. M., y Faulin, J. (2014). New opportunities in operations research to improve pork supply chain efficiency. *Annals of Operations Research*, *219*, 5–23. doi: doi:10.1007/s10479-013-1465-6

Rodríguez Sánchez, S. V., Pla-Aragones, L., y De Castro, R. (2018). Insights to optimise marketing decisions on pig-grower farms. *Animal Production Science*, *59* (6), 1126–1135. doi: doi:10.1071/AN17360

Schaeffer, S. E., Garza Villarreal, S. E., Espinosa Ceniceros, J. C., Sandra Cecilia, U. C., Nurmi, P., y Cruz-Reyes, L. (2018). A framework for informing consumers on the ecological impact of products at point of sale. *Behaviour & Information Technology*, *37*, 607–621. doi: doi:10.1080/0144929X.2018.1470254

Schaeffer, S. E., y Rodríguez Sánchez, S. V. (2020, enero). Forecasting client retention — a machine-learning approach. *Journal of Retailing and Consumer Services*, *52*, 101918. doi: doi:10.1016/j.jretconser.2019.101918

Xue, M., y Lai, C.-H. (2018). From time series analysis to a modified ordinary differential equation. *Journal of Algorithms & Computational Technology*, *12* (2), 85–90. doi: doi:10.1177/1748301817751480

Zhu, Y., Kumar, S., Rodríguez Sánchez, S. V., y Sriskandarajah, C. (2015, diciembre). Managing logistics in regional banknote supply chain under security concerns. *Production and Operations Management*, *24* (12), 1966–1983. doi: doi:10.1111/poms.12378

10. Justificación financiera de los requerimientos

El presupuesto es un total de $50,000 administrada en dos partes: la primera administración (mayo a agosto) consistirá en los rubros 107, 108 y 109 y la segunda administración (de septiembre a diciembre) consistirá en los rubros 101, 102 y 106.

101 Viáticos Apoyo para la presentación de los resultados del proyecto: alimentación y hospedaje del estudiante de doctorado al participar en un congreso para presentar su trabajo de tesis; se respetan los límites establecidos.

102 Pasajes Pago de transporte del estudiante de doctorado a un congreso para presentar resultados de su trabajo de tesis; el gasto ejercido entre los rubros de viáticos y pasajes corresponde al 30 % del monto apoyado, lo que es el máximo permitido.

103 Gastos de Trabajo de Campo *No se requiere.*

104 Ediciones e Impresiones *No se requiere.*

105 Servicios Externos *No se requiere.*

106 Cuotas de Inscripción Pago para la asistencia a congreso para que el estudiante de doctorado presente su trabajo.

107 Artículos, Materiales y útiles Diversos Compra de materiales para actividades diarias: papelería de oficina, consumibles, tóner de impresora.

108 Libros y Revistas Técnicas y Científicas Se contempla adquisición de libros según las necesidades de los alumnos tesistas.

109 Animales para Rancho y Granja *No se requiere.*

110 Becas Mil pesos mensuales a alumno tesista de licenciatura por los ocho meses del proyecto (mayo a diciembre), en un solo pago al finalizar la primera administración.

Se resumen los montos por rubro, indicando también el porcentaje del presupuesto total del proyecto.

Clave	Partida	Importe	Porcentaje
101	Viáticos	$7,000	15 %
102	Pasajes	$8,000	15 %
103	Gastos de Trabajo de Campo	—	0 %
104	Ediciones e Impresiones	—	0 %
105	Servicios Externos	—	0 %
106	Cuotas de Inscripción	$5,000	10 %
107	Artículos, Materiales y útiles Diversos	$5,000	10 %
108	Libros y Revistas Técnicas y Científicas	$17,000	24 %
109	Animales para Rancho y Granja	—	0 %
110	Becas	$8,000	16 %
Total	Gasto corriente	$50,000	100 %

11. Resultados beneficiados del PAICYT anterior

A partir del trabajo relacionado al proyecto IT512-15 que fue el último PAICYT de la responsable, se aplicaron las técnicas de pronóstico multifactorial en la tesis de maestría de J. A. Benavides Vázquez (2019) igual como en Schaeffer y Rodríguez Sánchez (2020).