Modeling the Spread of Air Pollution Across India Using Correlation Networks

Arya Maheshwari

In a Tortoiseshell: *In his Writing Seminar R3, Arya Maheshwari uses correlation networks to model air pollution data gathered in India. This excerpt, which is a condensed version of his introduction, demonstrates how quantitative papers can effectively utilize* ***global and scholarly motives*** *to communicate the importance of their high technical studies to a lay audience***.**

Excerpt

As pollutant emissions and large-scale fire events continue to rise globally with industrialization and climate change, combating air pollution for both the climate and human health has become a task of increasing importance worldwide, especially in regions like India, where a vast amount of the population lives in unsafe air conditions (Kamyotra, 2011; Agrawal, 2021). Air pollution has been shown to pose clear risks as a contributor to global deaths and mortality rates (Lim, 2012; Greenstone, 2015), and an extensive report conducted by IQAir on global air quality in 2020 found that 22 of the top 30 most polluted cities in the world were in India (IQ Air). These statistics demonstrate the clear need for stronger action in India, and the first step to combating air pollution through improved policy and efficient solutions is to effectively model its structure and spread.

In this study, we implement and examine a network-based computational model for air pollution dispersal and distribution, using a statistical approach based entirely on observed pollutant data that is relatively under-studied compared to widely-used machine learning methods (Bellinger, 2017), in order to better understand the spread of air pollution across India. The organization of particulate matter (PM) data by measurement stations allows for a natural and fruitful application of network science, namely through the use of time-series pollution data across Indian cities to generate correlation networks, a specific type of model previously suggested for dynamical systems in climate science (Yamasaki, 2008; Tsonis, 2006). While similar network-theoretic approaches have been employed in precursor pollution studies across regions of China and California (Wang, 2017; Dai 2017; Vlachogiannis, 2021), this approach, to the author's knowledge, has not been applied to Indian air pollution data: a search on Google

Scholar for "air pollution" and "correlation network" yields no studies focused on India, compared to the multiple results aimed at Chinese regions (over 10 verified). This gap thus presents an opportunity to use this method in a new region and derive novel, actionable results. Furthermore, in terms of analyzing methodology itself, this study demonstrates an application of a theoretical procedure to new real-world data, thereby allowing future researchers to better evaluate the potential of such an approach.

Concretely, the primary goals of this study are two-fold, based on the main strengths and capabilities of using networks as the fundamental computational tool. First, we aim to extract insights into the structure and spread of air pollution in India using network theoretic computations. Analyzing key global properties of our network enables an understanding of large-scale structure and trends of air pollution across India as a whole, while examining the properties of individual nodes provides information about specific cities and their effects on the broader system.

Second, this study also creates a product targeted towards lay users, with the key properties of being easy-to-use, concise, and visualizable. While existing quantitative models often achieve impressive predictive accuracy and numerical improvements, they are also harder to interpret and more resource-intensive to use, accessible primarily to academic researchers rather than users from the general public (Delavar, 2019; Xi, 2015; Srivastava, 2018). On the other hand, the underlying network representation of our computational model leads directly to intuitive visualizations of influence and similarity relationships between cities in terms of their air pollution. As such, in our approach we opt for giving users a product that enables qualitative understanding and direct visualizations, rather than highly technical models for numerical prediction.

Together, the two branches of this study lead to two major outcomes in terms of addressing the problem of air pollution. Insights from the computational model can be applied to generate more targeted, informed policy recommendations—at both national and local governmental scales—regarding how to most effectively combat air pollution. Furthermore, the user-end product fills the need for the public to gain an understanding of the influences and effects of air pollution in their own city, expanding the demographic to which research in air pollution modeling and insights about air pollution spread are directly accessible.

Author Commentary

Arya Maheshwari

As someone interested in quantitative disciplines like computer science, as well as policy-based solutions to broad problems in climate change currently facing the world, I wanted to find a topic for my R3 in WRI117 (Sustainable Futures) that would allow me to delve into the intersection of these areas. After some initial investigations into seminal modeling techniques in environmental science, I came upon the applications of correlation networks to climate research, which immediately caught my interest given some short previous forays into the world of networks and graphs. With correlation networks selected as my computational tool, I then looked into possible areas of application and decided on air pollution modeling as a relevant and important choice.

Thus, I arrived at what would ultimately be the focus of my paper, based half on technical methodology and half on relevant problems in climate change. Because of the quantitative nature of my project, the vast majority of my time was spent puzzling over ideas from statistics and network science and then developing the code that served as the backbone of my analysis. Despite the disproportionate amount of time I spent on wrangling code and understanding formulae, I learned throughout the process of drafting my essay that what you write about the most is not necessarily what you spent the most time working on. As such, while an overview of the technical aspects of the project was featured in the Methods section of my paper, I placed more emphasis while writing on thinking through my introduction and discussion sections to ensure that the work I was presenting was well-motivated and accessible. After all, any amount of quantitative work becomes irrelevant if there are no good reasons for and interpreted explanations of it.

Editor Commentary

Christina Cho

As a humanities major who has little-to-no knowledge of quantitative disciplines, I often struggle to fully comprehend the merit and significance of complex quantitative papers. However, I did not encounter this issue when reading Arya's paper. As this excerpt demonstrates, Arya adeptly incorporates several layers of **motive** that allow lay readers to fully grasp the global and scholarly significance of his highly mathematical study.

Arya opens his paper with a **global motive**, establishing the importance of his paper to "combating air pollution"—a highly global topic. While doing so, Arya also manages to explain why quantitative models are necessary when governments enact new policies. Arya then proceeds to introduce a **scholarly motive**, pointing out that his methodology utilizes a "relatively under-studied" set of data, which represents air pollution in India. In fact, as Arya writes, previous studies that involve correlation networks—Arya's selected methodology—have overlooked India altogether. Here, it's important to note how Arya cues his **scholarly motive** through specific words and phrases; he explicitly uses the word "gap," but the phrase "relatively under-studied" also points to a shortcoming in previous research. Arya further complicates his **scholarly motive** by highlighting the importance of applying his "theoretical procedure" to "real-world data." In other words, Arya is putting theory into practice, ultimately improving scholarly confidence in a particular methodology.

When discussing the two goals of his paper, Arya introduces yet another **scholarly motive**: In the past, highly quantitative studies have lacked "easy-to-use, concise, and visualizable" products for the general public. (Arya, of course, later provides examples of these products in his paper.) Given that this statement has implications for the greater public, we might also call it an example of **global motive**. Thus, Arya demonstrates how global and scholarly motives have the potential to intertwine.

If we take a step back to observe these **motives** altogether, we can see how Arya addresses multiple stakeholders with a single excerpt. His **global motive** addresses governments and all those impacted by climate change, while his **scholarly motive** addresses both scholars and lay individuals, who wish to better understand the issue of air pollution.

Professor Commentary

Andrea DiGiorgio, Princeton Writing Program

In this paper and excerpt Arya Maheshwari has found an important gap in both literature and application - how to measure and disseminate information about changes in air quality for the citizens of India. Specifically, he engages several layers of motive, moving from global (human health risks due to poor air quality), to scholarly (noting that a technological methodology has not been applied that could be very useful), to data (what are differences in the existing data from different methodologies that could inform this new methodology). His project is very ambitious in that it not only addresses this gap in knowledge and suggests a novel intervention, but also creates a proof-of-concept of this novel ambition that could (and hopefully will) evolve into an easily accessible product for the lay public to protect their health.

Works Cited

Anastasios A. Tsonis, Kyle L. Swanson, and Paul J. Roebber. 2006. What do networks have to do with climate? Bulletin of the American Meteorological Society 87, 5 (May 2006), 585–596. https://doi.org/10.1175/BAMS-87-5-585

Chavi Srivastava, Shyamli Singh, and Amit Prakash Singh. 2018. Estimation of Air Pollution in Delhi Using Machine Learning Techniques. In 2018 International Conference on Computing, Power and Communication Technologies (GUCON). IEEE, 304–309. https://doi.org/10.1109/GUCON.2018.8675022

Colin Bellinger, Mohomed Shazan Mohomed Jabbar, Osmar Zaïane, and Alvaro Osornio-Vargas. 2017. A systematic review of data mining and machine learning for air pollution epidemiology. BMC Public Health 17, 1 (Dec 2017), 907. https://doi.org/10.1186/s12889-017-4914-3

Dimitrios M. Vlachogiannis, Yanyan Xu, Ling Jin, and Marta C. González. 2021. Correlation networks of air particulate matter (PM2.5): a comparative study. Applied Network Science 6, 1 (Dec 2021), 32. https://doi.org/10.1007/s41109-021-00373-8

Girish Agrawal, Dinesh Mohan, and Hifzur Rahman. 2021. Ambient air pollution in selected small cities in India: Observed trends and future challenges. IATSS Research 45, 1 (Apr 2021), 19–30. https://doi.org/10.1016/j.iatssr.2021.03.004

IQAir. 2021. 2020 World Air Quality Report. https://www.iqair.com/world-most-polluted-cities/world-air-quality-report-2020-en.pdf

K. Yamasaki, A. Gozolchiani, and S. Havlin. 2008. Climate Networks around the Globe are Significantly Affected by El Niño. Physical Review Letters 100, 22 (Jun 2008), 228501. https://doi.org/10.1103/PhysRevLett.100.228501

Mahmoud Delavar, Amin Gholami, Gholam Shiran, Yousef Rashidi, Gholam Nakhaeizadeh, Kurt Fedra, and Smaeil Hatefi Afshar. 2019. A Novel Method for Improving Air Pollution Prediction Based on Machine Learning Approaches: A Case Study Applied to

the Capital City of Tehran. ISPRS International Journal of Geo-Information 8, 2 (Feb 2019), 99. https://doi.org/10.3390/ijgi8020099

Michael Greenstone, Janhavi Nilekani, Rohini Pande, Nicholas Ryan, Anant Sudarshan, and Anish Sugathan. 2015. Lower Pollution, Longer Lives: Life Expectancy Gains if India Reduced Particulate Matter Pollution. Economic and Political Weekly 50, 8 (2015), 40–46. http://www.jstor.org/stable/24481424

SJS Kamyotra, D Saha, SK Tyagi, AK Sen, RC Srivastava, and A Pathak. 2011. Guidelines for the Measurement of Ambient Air Pollutants.

Stephen S Lim et. al. 2012. A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. The Lancet 380, 9859 (Dec 2012), 2224–2260. https://doi.org/10.1016/S0140-6736(12)61766-8

Xia Xi, Zhao Wei, Rui Xiaoguang, Wang Yijie, Bai Xinxin, Yin Wenjun, and Don Jin. 2015. A comprehensive evaluation of air pollution prediction improvement by a machine learning method. In 2015 IEEE International Conference on Service Operations And Logistics, And Informatics (SOLI). IEEE, 176–181. https://doi.org/10.1109/SOLI.2015.7367615

Yue-Hua Dai and Wei-Xing Zhou. 2017. Temporal and spatial correlation patterns of air pollutants in Chinese cities. PLOS ONE 12, 8 (Aug 2017), e0182724. https://doi.org/10.1371/journal.pone.0182724

Yufang Wang, Haiyan Wang, Shuhua Chang, and Maoxing Liu. 2017. Higher-order network analysis of fine particulate matter (PM 2.5) transport in China at city level. Scientific Reports 7, 1 (Dec 2017), 13236. https://doi.org/10.1038/s41598-017-13614-7

Bios

Arya Maheshwari, 2025 is from Los Altos, CA and plans on majoring in either Mathematics or Computer Science. On campus, he is a part of the Naacho Dance Company and the Club Badminton team, and he enjoys pursuing his academic interests outside the classroom through clubs like Princeton ACM and research in quantum computing. In his free time, he is generally either out stargazing, on a walk, or eating comfort food. He wrote this essay as a first-year.

Christina Cho, 2024 is a Religion student also interested in Archaeology and East Asian Studies. She wrote this as a sophomore.