Constructing a Geospatial Data Pipeline for Environment and Health in South Korea – A Case Study of Personal Exposure to Fine Particulate Matter

Won Kyung Kim, Goeun Jung, Dongook Son, Kyumin Kim, Seokmin Ji, Soriul Kim, Chol Shin, Miji Kim, Hyunji Kim and Sun-Young Kim

ABSTRACT:

Geospatial data are powerful resources to understand geographic characteristics associated with environmental risk factors of human health. Diverse geospatial data can help identify specific geographic features responsible to air pollution and provide practical guidance to target such features in order to improve air quality and avoid adverse health effects. However, the significant heterogeneity in data sources and types as well as tremendous volume of data derived by temporal and/or spatial updates necessitate the geospatial data pipeline that comprises the series of steps including transferring raw data from disparate sources, modifying data, storing to a unified database, and constructing analytical platforms. Leveraging our recent work to construct a geospatial data pipeline in South Korea, this case study aimed to demonstrate the applicability of the data pipeline to public health research. Specifically, we examined the influence of various environmental characteristics on personal exposure to fine particulate matter ($PM_{2.5}$) in seniors.

Our geospatial data pipeline includes 8 categories of socio-demography, land cover, traffic, physical geography, vegetation, emissions, meteorology, and administrative boundaries. We obtained these data from various institutions such as Korea National Geographic Information Institute, Korea Transport Database, and U.S. National Aeronautics and Space Administration. After pre/post-processing of data, we computed geographic variables using the distance to features as well as the areas and proportions within specific buffer distances, such as the distances to the closest bus stop and the median normalized difference vegetation index within a 1,000-meter circular buffer. For a case study, our study population was 6 senior participants of the Korean Genome and Epidemiology and Korean Frailty and Aging Cohort Studies. Each participant carried a MicroPEM to measure real-time personal exposure to $PM_{2.5}$ with a GPS data logger for tracking the locations for 5 days in each of spring and summer. In addition, there were two other MicroPEMs installed inside and outside of the home for assessing indoor and outdoor $PM_{2.5}$. All $PM_{2.5}$ measurements collected every 10 seconds were gravimetrically corrected and averaged for 1 hour. The coordinates of all locations where participants visited and recorded in every second were converted to 1-hour mode. Using 355 geographic variables computed at 1,058 locations, we performed correlation analyses to select a subset of variables and regression analyses to examine the relationship between selected geographic variables and $PM_{2.5}$ concentrations by outdoor and personal exposure.

The regression analysis for outdoor $PM_{2.5}$ showed higher $PM_{2.5}$ concentrations at the locations with higher road lengths within a 1,000-meter and more registered vehicles in the surrounding area. When we applied the regression analysis to personal $PM_{2.5,}$ the increase in retail businesses and population in a 1,000-meter was associated with the increase in $PM_{2.5}$.

This research demonstrated the utility of a geospatial data pipeline to identify the environmental determinants responsible to particulate matter air pollution. Future research should expand the application of our geospatial data pipeline to diverse research of environmental challenges as well

as health endpoints, in order to enhance scientific knowledge of environmental risk factors and to contribute to developing solutions for fostering a sustainable environment.

KEYWORDS: *data pipeline, big data, spatial analysis, GPS, air pollution*

Won Kyung Kim, Department of Cancer AI & Digital Health, Graduate School of Cancer Science and Policy, National Cancer Center, Goyang-si, South Korea

Goeun Jung, Department of Cancer AI & Digital Health,

Graduate School of Cancer Science and Policy, National Cancer Center, Goyang-si, South Korea

Dongook Son, Department of Cancer AI & Digital Health, Graduate School of Cancer Science and Policy, National Cancer Center, Goyang-si, South Korea

Kyumin Kim, Department of Cancer AI & Digital Health, Graduate School of Cancer Science and Policy, National Cancer Center, Goyang-si, South Korea

Seokmin Ji, Department of Public Health Science, Graduate School, Korea University, Seoul, South Korea

Soriul Kim, Institute of Human Genomic Study, College of Medicine, Korea University, Seoul, South Korea

Chol Shin, Institute of Human Genomic Study, College of Medicine, Korea University, Seoul, South Korea

Miji Kim, Department of Biomedical Science and Technology, Graduate School, Kyung Hee University, Seoul, South Korea

Hyunji Kim, Department of Cancer AI & Digital Health, Graduate School of Cancer Science and Policy, National Cancer Center, Goyang-si, South Korea

Sun-Young Kim, Department of Cancer AI & Digital Health, Graduate School of Cancer Science and Policy, National Cancer Center, Goyang-si, South Korea