Geospatial Big Data Fusion: A Study of the Effects of Renewable and Non-renewable Resources in Los Angeles County

Ryan Huppert
Army Educational Outreach Program
University of Nevada, Las Vegas
Las Vegas, USA
ryan.huppert01@gmail.com

Lyra Dema
Army Educational Outreach Program
University of Nevada, Las Vegas
Las Vegas, USA
lyradema03@gmail.com

Kaiyan Wilson
Army Educational Outreach Program
University of Nevada, Las Vegas
Las Vegas, USA
kaiwilson412@gmail.com

Justin Le
Army Educational Outreach Program
University of Nevada, Las Vegas
Las Vegas, USA
justin.le1290144@gmail.com

Taryn Thompson
Research Experiences for Teachers
University of Nevada, Las Vegas
Las Vegas, USA
thomptn@nv.ccsd.net

David Garcia
Research Experiences for Teachers
University of Nevada, Las Vegas
Las Vegas, USA
garcid13@nv.ccsd.net

Andrea Sedano
Research & Engineering Program
University of Nevada, Las Vegas
Las Vegas, USA
sedanoavs@gmail.com

Jacob Howard
Academic Scholar
University of Nevada, Las Vegas
Las Vegas, USA
jhoward@unlv.nevada.edu

Hadi Salman
Computer Science Department
University of Nevada, Las Vegas
Las Vegas, USA
hadi.salman@unlv.edu

Justin Zhan
Computer Science Department
University of Nevada, Las Vegas
Las Vegas, USA
justin.zhan@unlv.edu

Laxmi Gewali
Computer Science Department
University of Nevada, Las Vegas
Las Vegas, USA
laxmi.gewali@unlv.edu

Paul Oh
Mechanical Engineering Department
University of Nevada, Las Vegas
Las Vegas, USA
paul.oh@unlv.edu

***Abstract*— Air pollution has a direct correlation to population and is a prevalent crisis throughout various societies. Air quality has deteriorated globally due to the rise of a multitude of pollutants, directly affecting living conditions for people who come into contact with them. To improve efficiency of data analysis and therefore increase awareness of the crisis, researchers can use geospatial data with data interpolations to predict pollution in metropolitan areas. This application will help address health and environmental concerns. The objective of this research is to use geospatial data analysis and satellite imaging to create a visual representation overlay of pollution on a geospatial imagery.**

***Keywords*— Data Fusion, Big Data, Environmental Monitoring, Air Pollution, Urban Areas, Predictive Modelling**

I. INTRODUCTION

Pollution and greenhouse gasses have increased across the United States during the past several decades. Air pollution can cause many health conditions, including asthma, heart disease, and low birth weight in infants [1]. For example, California has developed the "worst US air pollution levels" [2]. California's response to all of the pollution and greenhouse gasses was to pass AB 32, the California Global Warming Solutions Act, also known as the Cap and Trade program. AB 32 requires a reduction of greenhouse gas emissions to 1990 levels by 2020 and sets the stage for the transition to a sustainable low-carbon future. AB 32 takes a comprehensive approach to improve the environment and natural resources while maintaining a strong economy [3]. With big data fusion, researchers can combine different data sources (such as population, pollution, non-renewable power plants, and renewable power plants) to visualize the effect of renewable energy on the pollution levels. A geospatial imagery layout is utilized to help display the data in a user-friendly interface. The application can also be used as a comparison between regions to visualize the differences between states that are renewable energy dependent or non-renewable dependent. Consequently, this application could also be used to increase awareness of air pollution and its harsh effects on people's health. For the purpose of this study, California was chosen for renewable resources and Kentucky for non-renewable resources.

II. RELATED WORK

A. Geospatial Data

Geospatial data fusion creates new associations involved with processing and abstraction and creates a new data element through observation, object/feature, and decision fusion [4]. The implementation of new deployment platforms and sensor types increases the variety and complexity of geospatial observation. Such implications and developments within big data technology assemble a collection of easily exchangeable information within remote sensing research, consistent with geospatial volumes in order to enhance data analysis [5].

In response, a live data feed is created with a responsive server-side application involved with real-city application and several other data importing operations [6]. The increasing variety and volume of geospatial data demand greater abilities

to combine and associate information from multiple sources to create knowledge about the geographic world. Through the Geographic Information Systems (GIS), the implementation of the geographic and big data environments allows researchers to plot data points with the support of sensor networks and data analysis [7].

B. Big Data Fusion

The information trends resulting from the OGC Standards baseline linked data, and other sources of geospatial observations have been used in order to increase the variety and volume of geospatial data [8].

In principle, a decentralized data fusion system is more challenging to implement because of the computation and communication requirements. However, in practice, there is no single best architecture, and the selection of the most appropriate architecture should be made depending on the requirements, demand, existing networks, data availability, node processing capabilities, and organization of the data fusion system.[9]

The integration of data and knowledge from several sources is known as data fusion. In general, all tasks that demand any parameter estimation from multiple sources can benefit from the use of data/information fusion methods. For the purpose of this article, just going to is data fusion. [10]. The most agreed upon definition of data fusion was provided by the Joint Directors of Laboratories (JDL) [11]: *"A multi-level process dealing with the association, correlation, combination of data and information from single and multiple sources to achieve refined position, identify estimates and complete and timely assessments of situations, threats, and their significance."*

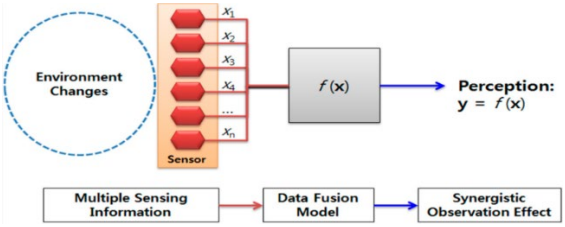

Fig. 1. This is a figure by (H. Kim and D. Suh, "Hybrid Particle Swarm Optimization for Multi-Sensor Data Fusion," *MDPI*, 24-Aug-2018. [Online]. Available: https://www.mdpi.com/1424-8220/18/9/2792. [Accessed: 29-Jun-2019).

C. Global/Environmental Challenges

Air pollution is a negative phenomenon that affects not just humans, but the environment as well. Air pollution, in particular, has accelerated climate change, created an ozone hole, and increased particulate matter in the air. The respiratory, cardiovascular, nervous, and immune systems of humans can be damaged by pollution [2]. Some progress has been made to prevent future harm by prohibiting the use of some particularly harmful chemicals and by forecasting air pollution. Several AI systems can forecast the pollution much better than statistical measures, but there is not a direct comparison between all of them [3], [7].

With the rise of cities and the growth of the population, air pollution has become more dangerous than before and continues to threaten the health of the general population. The mortality rate in heavily polluted areas leads to very high mortality rates, particularly in developing countries. A figure known as the Environmental Kuznets Curve displays the relationship between industrialization and air pollutants - As a country begins to develop economically, a variety of factors will create monumental amounts of pollution if safety precautions are not taken. Some of the claimed effects of air pollution cannot yet be traced to it. However, the number of people afflicted, as well as the data of deaths in areas of high pollution, gives support to the dangers of air pollution [10].

D. Energy Studies

Renewable energy in California is climbing as strides against climate change. The power grid in California to date is mostly comprised of renewable resources including photovoltaic, solar, wind, and biomass. PV (photovoltaic) and wind energy, in particular, have grown in California, which makes up around 50% of the renewable energy going to the grid [13].

California has signed many bills in the Clean Energy & Pollution Reduction Act, setting goals to have renewable energy encompass 33% of energy in the state by 2020, and 60% (formerly 50%) by 2030. As of November of 2018, the 33% goal has already been surpassed, with California even constructing homes with solar panels. The Million Solar Roofs Initiatives set in 2006 set a goal of a million solar roofs by 2018. While California did not quite reach that goal, they were very close, with about 958,000 constructed. The push in California for solar power is so significant that financial aid was given to residents that chose to invest in solar panels [14]. The observation of data in air pollution from non-renewable to renewable can show the effectiveness of the transition.

E. Predictive Modeling

Satellite rendering and observation with the application of machine learning algorithms are more prominent in predictive learning/modeling to enhance the accuracy of PM2.5, an atmospheric component's prediction. Uneven spatial coverage and point-based monitoring work hand in hand with the analysis of health effects, epidemiology, and climate effects to aid in the study of atmospheric pressure and concentrations. New advancements as such further contribute to pollution forecasting as it helps understand the fatality of climate change and the increasing danger of atmospheric shifts [14].

Air quality forecasting and modeling revolve around machine-learning predictive algorithms, specifically the use of accurate sensor readings as well as complex calculations in order to analyze and predict the current quality index of the atmosphere. Data-driven information used to analyze and predict air pollution risks requires the incorporation of gas sensors, most accurately requiring the integration of neural network analysis to understand environmental patterns that could be disrupted by pollution; most of such neural network predictive models ranging accuracy from 94.2-99.6% [15].

In relation to our current experiment; however, data analysis involves simpler algorithms in order to display more regressive models for observation. Spatial mapping of air pollution involving the systematic recording, collection, and archiving of the meteorological elements of the studied area. These findings combine into a predictive system more fitted to the local environment in order to increase predictive accuracy. Such advancements become prominent primarily in operational application of atmospheric analysis [16].

III. Implementation/Methodolgy

A. Data Representation

The big data we are analyzing come from several sources, including Electricity Generation Data from the CARB Pollution Mapping Tool on the California Air Resources Board website [3] and Solar Measured Production Data from the California Solar Initiative Data on the Go Solar California website [17]. The CARB Pollution Mapping Tool provides statistical data on Cement Plants, Hydrogen Plants, Oil and Gas Production, Cogeneration, Refineries, and Power Plant Emissions. The focus of this research is the harmful effects of power plant emissions. Therefore, the yearly data regarding power plant emissions was interpolated into a dataset identified as Electricity Generation Data. The Electricity Generation Dataset created includes the type of plant, the addresses of power plants that are a part of Cap-and-Trade in California, the amounts of Greenhouse Gases produced by each power plant yearly, and locations of each plant by county. Our research requires the location of power plants and the amount of CO2 produced per year. The location data is interpolated from addresses into latitude and longitude of the power plants. The CO2 data was utilized because it is the most prevalent greenhouse gas and provides an accurate representation of all greenhouse gases. The data acquired from the Solar Measured Production Data includes zip code, production period end date by month, and period kWh production. The data is monthly and was transferred into an annual format numbers with the Period kWh Production Data Fusion.

B. Data Fusion

After removing irrelevant information, we meshed the datasets into a single set through data fusion. It was fused with an application called C# (C sharp). The data was then combined to create meaningful information through extensive geospatial data analysis.

The application and analysis of data sets from differing domains harmonize contradicting perspectives and studies in order to provide in-depth results and solutions to the issues. Data fusion combines the findings from geospatial mapping, predictive modeling, and environmental studies help support modes of change. Such efforts are beneficial to multiple parties dependent on the circumstances within the modeled areas.

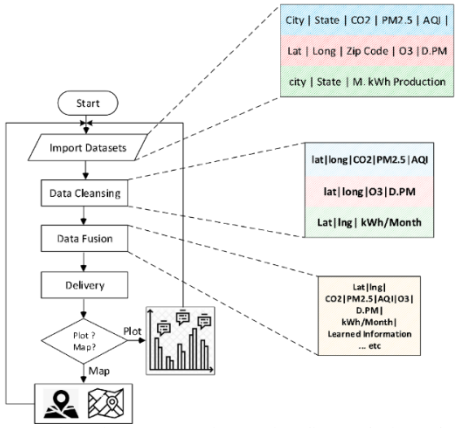

Fig. 2. This diagram explains the flow of data fusion and how we implemented our data.

C. Data Delivery

The demonstration application prompts the user to upload four distinct datasets to visualize the data. Through importing our sample data, we are given the option to choose between plotting map data and displaying data in a chart form. Our example utilized small, controlled environments in both California and Kentucky to show comparable features between unique locations within our software. When choosing to plot map data, the model plots markers across California and Kentucky to represent the location and area-based sections in both states. Each of these markers allows the user to view the air quality in a given area by hovering via cursor. The charting option displays a 2-variable histogram representing various renewable and non-renewable parameters to be compared between states. California is represented in blue while Kentucky is green to differentiate between states to showcase contrasts to make it easier for users to compare the parameters within the given datasets.

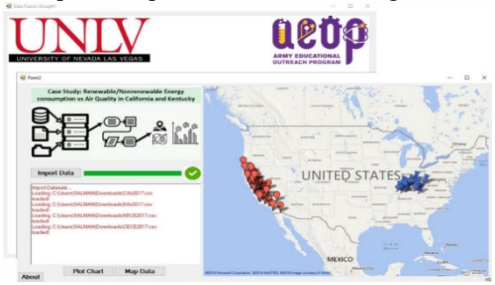

Fig. 3. This is a capture imagery of our demonstration.

IV. Results

The application shows us that Kentucky primarily uses non-renewable resources (coal) to power their energy plants. On the other hand, California's usage of non-renewable resources (coal) for energy production is substantially lower. California uses different types of renewable resources to power its energy grid that includes natural gas, ethanol, hydroelectric, and biomass. Once we compare the air quality between California and Kentucky, the results show that Kentucky has a higher PM 2.5 (Fine Particulate Matter) level. As a result, the air quality from Kentucky is much worse. As well as the death rates caused by lung cancer and brocades this in Kentucky and California.

V. Analysis and Discussion

Big Data Fusion is a revolutionary field with many future applications that can be explored by forthcoming researchers. This research explores several tools and methods that were used through Data Representation, Data Fusion, and Data Delivery. The data sets included information on how each state produces energy and also on the air quality is at every state down to the local level. Fusing the data, we can observe that states who are dependent on coal for energy have a higher PM 2.5 (Fine Particulate Matter) particulate in the atmosphere. The data also included energy sources to produce electric power includes coal, natural gas, ethanol, hydroelectric power, biomass, and other renewables. Using C-sharp, the data was plotted along California and Kentucky using the Longitude and Latitude coordinates. The application used C-sharp to allow the user to input data that can be overlaid on a map or demonstrated on a graph. Then the user can hover over and read the air quality level as

measured by the PM 2.5 level. States like California that have a lower PM 2.5 level demonstrate better air quality.

Future implementations of our findings could potentially generate data with other types of pollution threatening both our atmosphere and wildlife; specifically, oceanic territories. In-depth analysis of waste in major bodies of water with the use of our code to detect its abundance in certain areas. The use of our predictive models could configure means of pollution regulations and limitations, as well as extensive studies of oceanic wildlife and the negative effects of waste on the native species/habitats.

ACKNOWLEDGMENT

The researchers would like to acknowledge Dr. Justin Zhan, Hadi Salman, the Army Education Outreach Program (AEOP), Research & Engineering Apprenticeship Program (REAP), the National Science Foundation, University of Nevada, Las Vegas, and Research Experiences for Teachers (RET). The publication is a direct outcome of a collaboration between these programs as they allowed secondary students and teachers the opportunity to accomplish graduate-level research, working with graduate students and professors to pursue a solution to a university-level research project

REFERENCES

[1] staff, “California has worst US air pollution: report,” Phys.org,19-Apr-2018.[Online].Available:https://phys.org/news/2018-04-california-worst-air-pollution.html. [Accessed: 14-Jun-2019].

[2] oehha.ca.gov.[Online].Availabl:https://oehha.ca.gov/calenviroscreen/indicators. [Accessed: 14-Jun-2019].

[3] California Air Resources Board, “Assembly Bill 32 Overview,” California Environmental Protection Agency Air Resources Board. [Online]. Available: https://www.arb.ca.gov/cc/ab32/ab32.htm. [Accessed: 14-Jun-2019].

[4] J. Geng, S. Wang, W. Gan, H. Yuan, Z. Chen, and T. Dai, “Promoting Geospatial Service from Information to Knowledge with Spatiotemporal Semantics,” Complexity, 21-Jan-2019. [Online]. Available: https://www.hindawi.com/journals/complexity/2019/9301420/. [Accessed: 25-Jun-2019].

[5] G. Percivall and T. Taylor, "Advances in fusion of big geospatial data," 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, 2017, pp. 380-383.

[6] P. A. Parikh and T. D. Nielsen, "Transforming traditional geographic information system to support smart distribution systems," 2009 IEEE/PES Power Systems Conference and Exposition, Seattle, WA, 2009, pp. 1-4.doi: 10.1109/PSCE.2009.4839979

[7] Chen, B., & Kan, H. (2008). Air pollution and population health: a global challenge. Environmental health and preventive medicine, 13(2), 94–101

[8] . R. E. Sorace, V. S. Reinhardt, and S. A. Vaughn, “High-speed digital-to-RF converter,” U.S. Patent 5 668 842, Sept. 16, 1997.

[9] M. Shell. (2002) IEEEtran homepage on CTAN. [Online]. Available: http://www.ctan.org/tex-archive/macros/latex/contrib/supported/IEEEtran/

[10] J. Zhang, J. Jorgenson, T. Markel and K. Walkowicz, "Value to the Grid From Managed Charging Based on California's High Renewables Study," in IEEE Transactions on Power Systems, vol. 34, no. 2, pp. 831-840, March 2019.

[11] J. Roy, “From Data Fusion to Situation Analysis,” https://www.semanticscholar.org/paper/From-Data-Fusion-to-Situation-Analysis-Roy/ff790d907dc0f53d6c90342f7ec90053dfae0827?citationIntent=methodology#citing-papers, 2001. [Online]. Available: http://fusion.isif.org/proceedings/fusion01CD/fusion/searchengine/pdf/ThC21.pdf. [Accessed: 21-Jun-2019]

[12] Kethireddy, S. R., Tchounwou, P. B., Ahmad, H. A., Yerramilli, A., & Young, J. H. (2014). Geospatial Interpolation and Mapping of Tropospheric Ozone Pollution Using Geostatistics. International journal of environmental research and public health, 11(1), 983–1000.

[13] Partain, Larry, and Lewis Fraas. “Displacing California's Coal and Nuclear Generation with Solar PV and Wind by 2022 Using Vehicle-to-Grid Energy Storage.” 2015 IEEE 42nd Photovoltaic Specialist Conference (PVSC), 14 June 2015

[14] California Energy Commission - Tracking Progress Appendix M. Weng-GutierrezEnergy Commission, “California Energy Commission – Tracking Progress,” California Energy Commission Tracking Progress, 10-Jan-2019.

[15] Bai, L., Wang, J., Ma, X., & Lu, H. (2018). Air Pollution Forecasts: An Overview. International journal of Bellinger, C., Mohomed Jabbar, M. S., Zaïane, O., & Osornio-Vargas, A. (2017). A systematic review of data mining and machine learning for air pollution epidemiology. BMC public health, 17(1), 907.

[16] T. M. Amado and J. C. Dela Cruz, "Development of Machine Learning-based Predictive Models for Air Quality Monitoring and Characterization," TENCON 2018 - 2018 IEEE Region 10 Conference, Jeju, Korea (South), 2018, pp. 0668-0672.

[17] Silas Michaelides, Dimitris Paronis, Adrianos Retalis, and Filippos Tymvios, “Monitoring and Forecasting Air Pollution Levels by Exploiting Satellite, Ground-Based, and Synoptic Data, Elaborated with Regression Models,” Advances in Meteorology, vol. 2017, Article ID 2954010, 17 pages, 2017.

[18] Bai, L., Wang, J., Ma, X., & Lu, H. (2018). Air Pollution Forecasts: An Overview. International journal ofBellinger, C., Mohomed Jabbar, M. S., Zaïane, O., & Osornio-Vargas, A. (2017). A systematic review of data mining and machine learning for air pollution epidemiology. BMC public health, 17(1), 907.

[19] T. M. Amado and J. C. Dela Cruz, "Development of Machine Learning-based Predictive Models for Air Quality and Characterization," TENCON 2018 - 2018 IEEE Region 10 Conference, Jeju, Korea (South), 2018, pp. 0668-0672.

[20] *California Solar Statistics*. [Online]. Available: https://www.californiasolarstatistics.ca.gov/data_downloads/. [Accessed: 26-Jun-2019].