IJNRD.ORG ISSN : 2456-4184

INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG

An International Open Access, Peer-reviewed, Refereed Journal

A Review on GeoSpatial Artificial Intelligence Techniques to Handle Missing Data

Nukala Naga Sai Venkat, Karthik Chitteti, Oliver Isaiah, Koushik L

Student, Student, Student, Student

dept. of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation Vaddeswaram, Andhra Pradesh, India

Abstract: Geospatial Artificial Intelligence (GeoAI) is emerging to top of its game with its wide range of applications. In this paper we are going to review about the present trends in GeoAI, machine technologies and the most common challenges that we come across when using the temporal data for predicting the Environmental Epidemiologies (EE) most precisely in air quality prediction. Geospatial Artificial Intelligence (GeoAI), which is an emerging scientific discipline, combines the innovations in spatial science, deep learning, data mining and temporal data to extract knowledge from spatial big data. Spatial data and spatial science plays a huge role in understanding and visualizing the real world phenomenon based on their locations. Spatial scientists use geographical Information Systems (GIS) to achieve patterns in the spatial big data. These GeoAI technologies provide an important advantage for exposure modeling in environmental epidemiology.

Index Terms **- Geospatial Artificial Intelligence(geoAI), Environmental Epidemiology (EE), Geographical Information Systems (GIS), Artificial Intelligence, Machine Learning (ML)**

INTRODUCTION

Artificial Intelligence and machine learning methods has been increasingly used in the recent times in weather forecasting, health, healthcare and many other domains. It is also used in predicting the diseases and in understanding the climatic changes and their effect on the human life [*]. The study of the effect on human health due to the physical, biological and chemical changes in the environment in called as environmental epidemiology (EE). GeoAI contributes very a lot to the EE in terms of understanding the trends and exposures to the human health. GIS is a technology that integrates geographical science to the data for collaboration and understanding. GIS uses all types of data including spatial data and images. Remote sensing (RS) is another technology which is useful in collecting the data that can be analyzed using GIS and GeoAI techniques for predicting the EE.

RESEARCH METHODOLOGY

For understanding the present technology and its challenges, study is needed. After examining the work from various sources many observations were done.

J. Alper et al., proposed one of the methods to predict the chemical toxicity in the environment. The proposed method requires us to make a large neural network having all the compounds from the dataset. The data need to be of high quality as there are different chemical components with a similar structure [1].

Weichenthal et al., proposed the use of convolutional neural networks for predicting the environmental exposures. Predicting the air pollution in a city could be possible with just the images being fed to the model. If the image is of lesser quality, then wrong predictions could be made [2].

Gonzales et al., suggests about the major role GeoAI plays in making models for hydraulic and hydrological modelling. They not only but also discussed the major flaws GeoAI possess right now, and a few suggestions about the way to solve them [3]

Vo Pham et al., proposed all the potential applications of GeoAI and future applications are listed in their work. The use of exposure modelling is mentioned. GeoAI's capability of handling big data is discussed [4].

Bhaskaran et al., proposed on a specific approach of time series regression. This technique is used to predict the ozone concentration of the atmosphere. As the data increases, the prediction will get more accurate, but the data size would increase [5].

Neelakandan et al., has developed a new ETAPM-AIT method for predicting air quality using IoT and ML techniques. At the initial stage, IoT-based sensors are used to collect details related to pollutants. Hardware must be maintained well in order to get the best results [6].

Chauhan et al., shows us about how different countries are taking up the GeoAI and the way of approach each country must develop. Although all the countries are showing good contributions for the development, there are still some problems like data insufficiency, data inaccuracy and such. Much more changes are to be made [7].

EHRTC et al., focused on how to handle the data we use for prediction in our GeoAI. In specific, pointed on the use of Bayesian Networks, and their future prospects of usage in GeoAI [8].

Temenos et al., focused on using XAI (Explainable AI) and Remote Sensing in order to predict the number of COVID-19 cases and deaths. It also implemented Random Forest technique and there were mixed results. The spread and mortality rate are too spread out to make a fixed conclusion [9].

Schmidt et al., offers a simple introduction on how to use AI techniques to apply on the environmental factors. Pretty simple and straight forward examples are used. They are just elementary level predictions and cannot be taken seriously [10].

Pearce et al., focuses on how to use a technique called casual inference in GeoAI to predict results. Now, casual inference is a method to handle imperfect data, so it is one of the ways to combat a data problem. But in the end, assumption is not truth. So, there could be slight deviation to the actual result [11].

GREGORY et al., focussed on the wild bird population in a particular area, and their derivations to the environmental health of that area. Wild birds are the best indicators of the health of an area. But as the farmlands are getting modernized, and habitats destroyed. It is becoming difficult to use them as indicators, also there are various bird species that act differently [12].

Boulos et al., uses the big data that is collected through all sorts of activities, and the data was used to predict the health and well being of people at a particular area. Now as the data increases, so does the accuracy [13].

Eaker et al., discussed the problems that occur when imperfect data is used. Data organizing is a key point to hold when we use data. When the volume of data increases, it also means that organizing data at that point will be difficult. So, it is advised to organize data at an early stage such that no problem occurs [14].

McFarland et al., proposed a solution for handling the big data by segmenting the data into smaller bits. By doing so it will be easier to handle the data [15].

Mezias et al., tells us about the need to carry on, even if the data is imperfect and not complete. Assumptions and guesswork are the best ways to handle the data. But by doing that, we may be making a fatal mistake, or our assumption could be correct, and it would mean success [16].

Browna et al., discussed the need for removing the errors and how they can be fatal to the model [17].

Azim et al., suggests the use of the cloud service infrastructures available, like IaaS, PaaS and such. This paper introduces the DaaS, which is used to handle data that is continuously updated. DaaS processes raw data into processable data that could be used in AI techniques [18].

Vandermeer et al., proposed a way to use discrete mathematic techniques to handle the continuously updated data. For a small amount of data, it will be sufficient. But for big data, using techniques like mean difference and standardized mean difference, are heavily CPU intensive tasks [19].

Tripepi et al., proposed a way to handle the big data which is specifically used to epidemiological questions. Many methods such as multiple linear regression and multiple logistic regression are used. These processes are lengthy and intensive but give good results [21].

Katharina et al., discussed the traditional and new recent approaches on how to handle continuously updated data. Many ML methods are being used to handle the data. The famous ways include time series regression, power analysis and such. While the new methods are viable, they are not exactly accurate and precise. But as new methods evolve, the handling of data could be even more perfected [22].

Author and Year published	**Proposed Solution**	**Challenges**	**Advantages**	**Limitations**	**Results**
J. Alper et al.,(2019)	Make a large neural network having all the compounds using the dataset.	To use AI and ML techniques to predict the chemical toxicity.	By using the AI model, we can get the accurate prediction	high quality data, or predictions will be wrong	Using AI and ML, we can make good advances in regarding to research about environmental health.
Weichenthal et al., (2018)	We use deep convolutional neural networks to estimate environmental exposures	Use the images to find the concentrations of pollutants in air using deep learning techniques	We will be able to find the amount of pollution in a city just by sending some images to our model.	The images should be of good quality. Or prediction will go wrong.	complement existing measurements for data-rich settings and could enhance the resolution and accuracy ofestimates in data poor scenarios.
Gonzales et al.,(2022)	machine learning (ML) and parallel computing- spatial and non-spatial dataset effectively	GeoAI and machine learning applications in hydrological and hydraulic modelling.	non-linear modeling, computational efficiency, integration of multiple data sources, high accurate prediction capability.	adequate model setting and low physical interpretability, explainability, and model generalization.	Hydrological GeoAI -integrating the physical-based models' principles with the GeoAI - autonomous prediction and forecasting systems.
Vo Pham et al.,(2018)	Exposure modelling exposure assessment to determine the distribution of exposures in study populations.	To use the big data of heavy volumes available	exposure modeling in environmental epidemiology, computational efficiency, flexibility in algorithms and workflows to accommodate relevant	The data we use for training the model should be of high qualityand all the variables of correct accuracy.	provide an overview of key concepts surrounding the evolving and interdisciplinary field of geoAI including spatial data science, machine learning, deep learning, and data mining.
Bhaskaran et al.,(2013)	Use time series data on the exposure of ozone gas vs no. of deaths for the model using time series regression	To find the changes in concentrations of ozone gas in the atmosphere.	The model will be giving accurate results at any time as we can give it data continuously.	The amount of data will be increased regularly.	plotting and tabulating the data, controlling for confounding, presenting exposure effects appropriately and model checking
Neelakandan et al.,(2021)	The proposed ETAPM-AIT model includes a set of IoT based sen- sor array to sense eight pollutants namely NH4 , CO2, NO2 , CH4 , CO , PM , temperature and humidity.	To detect air pollution using AI and ML techniques with IoT.	There will be continuous monitoring of air pollutants, by using a technique air quality index (AQI) as there will be sensors present.	As there is hardware present, there will be a chance of wear and tear, and need to be constantly maintained.	This paper has developed a new ETAPM-AIT method for predicting air quality using IoT and ML techniques. At the initial stage, IoT-based sensors are used to collect details related to 8 pollutants.
Chauhan et al.,(2021)	How various countries around the globe make their research regarding the best way to use GeoAI.	Insights about the latest trends in GeoAI in countries around the world.	Many countries show good progress in development and handling of data and such.	There are still problems like insufficient data, inaccurate data and such.	The development of GeoAI is progressing slowly but, there are some challenges to be solved.
EHRTC et al.,(2018)	This paper particularly focuses on Bayesian networks and their potential applications for GeoAI.	This paper shows the various ways to handle the data we receive.	Bayesian networks are a good technique to handle the GeoAI data we receive.	Making networks from data for Bayesian data is difficult.	Using Bayesian networks for GeoAI is like a double edged sword.
Temenos et al.,(2022)	Using Random Forest technique, to make a model on Covid-19 case/death predictions.	How to predict the spread of Covid-19 virus using AI methodologies.	One major find is relation between the Ozone concentration and spread of the virus.	The predictions we got are not completely reliable, there are many loopholes.	The spread and mortality rates of countries is very widely distributed.
Schmidt et al.,(2020)	Basic Machine Learning techniques are used to simple data, for elementary level predictions.	Basic Knowledge necessary to dwell deep into AI and ML	Easy to execute and anyone could be able to do it.	The process is too primitive and cannot be used seriously.	There are many possible ways of implementing as there are the applications of GeoAI.
Pearce et al.,(2020)	Causal inference is a data control technique and a way to handle imperfect data, which is well suited for the data of GeoAI.	Causal inference is the process of determining the independent, actual effect of a particular phenomenon that is a component of a larger system.	We can be able to fill in all the holes in the data and could train the models more effectively.	In the end, they are just assumptions, which could completely mislead the model.	All methods have assumptions that are often not possible to (fully) test.
GREGORY et al.	An obvious measure to focus on, when understanding the state of biodiversity and how it is changed.	To use the wild birds population trends to determine the environment quality.	There are a number of reasons to think that birds as a group might act as reasonable biodiversity indicators.	The degree to which a single taxon can faithfully represent the status and trends in other taxa is a matter of debate.	As there is habitat loss, due to many factors, The bird species are getting disturbed and the data acquired is also inaccurate.
Boulos et al.	GeoAI is an currently in development AI technology but has a good scope in future applications.	This paper covers the various outcomes that can be achieved using GeoAI.	There are many ways the humanity could be benefited by using GeoAI in various fields.	As massive amounts of data continue to be captured and collected, privacy could be problematic.	There is an emerging role for GeoAI in health and healthcare as location is an integral part of both population and individual health.

CHALLENGES

The most common challenges of using temporal data are inaccurate or missing data. And one of the biggest challenges is the continuously changing data.

A. **Inaccurate data:** Inaccurate data refers to the data with missing values or the data that needs imputation. This data impacts the working of the model. It is very important to use high quality data to get accurate results.

B. **Dynamic data**: Dynamic data refers to data that keeps on changing daily such as the weather, which is one of the most crucial data for environmental epidemiological studies. It is very important to handle such data.

Geospatial Artificial Intelligence(geoAI) has seen many advancements in the recent past. These advancements paved way for the use of geoAI techniques in understanding the environmental epidemiology (EE) effectively. To determine the main factors and exposures that influence the outcome of our model we will be using exposure modeling and exposure assessments. Exposure modeling involves the development of a model to represent a particular environmental variable using various data inputs such as environmental measurements and statistical methods such as land using regression and generalized additive mixed models. Exposure modeling is a cost-effective approach to assess the distribution of exposures in particularly large study populations compared to applying direct methods. This method uses geospatial temporal data for understanding the exposures geoAi requires large spatial data for predicting the environmental epidemiology. Artificial Intelligence and Deep learning techniques such as timeseries forecasting and random forest are used for predicting the EE. These methods require data without any missing values. Missing values can be handled using various methods.

TECNIQUES TO HANDLE MISSING DATA:

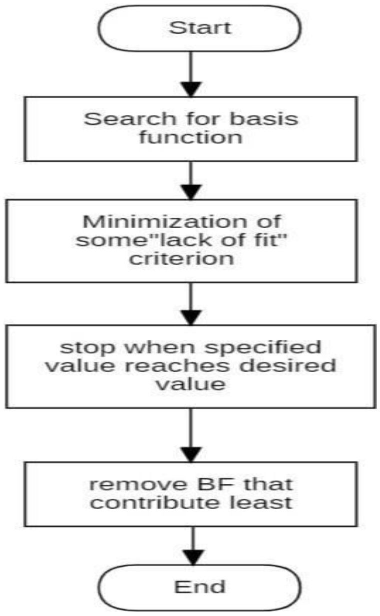
Techniques to Handle Missing Data The major challenges in using the geospatial temporal data are noisy and missing data. Handling such data is the most important step in designing the model. It is crucial to clean and process the data for modelling. The missing data can be handled using various techniques such as:

A. Statistical methods:

Missing data can be handled using the standard statistical imputations such as mean, median, mode. But this method is suitable for small datasets but in case of larger datasets, this method is not very useful. Whereas for larger datasets a binning technique is used. Binning involves dividing the dataset into small bins or clusters and the missing attributes are imputed with the mean, median or mode. This way the dataset retains most of its observations and gives better results.

B. MARS (Multivariate Adaptive Regression Splines):

MARS belongs to a group of regression algorithms which are used to predict continuous data. It is an algorithm for complex non-linear regression problems. The MARS algorithm builds multiple linear regression models across the range of predictor values. It involves partitioning of data and running a linear regression on each part.
1) Forward Stage: In this stage the algorithm creates a collection of basis-functions (BF) and the range of predictor values are partitioned into several groups. Then a separate linear regression is modelled for each such group. These regression lines are then connected. The connections are called knots. The algorithm automatically searches for the best spots to place the knots.
2) Backward Stage: MARS estimates a least-squared model with BF as independent variables. It fits a very large model which is subsequently pruned to avoid overfitting. It is done in an iterative manner by removing the BF that contributes the least to model fit.


Start
Search for basis function
Minimization of some"lack of fit" criterion
stop when specified value reaches desired value
remove BF that contribute least
End

C. VAR (Vector Auto Regression):

Vector Autoregression (VAR) is a multivariate forecasting algorithm that is used when two or more time series influence each other. For this method to be used the dataset needs to have at least two time series variables and the time series should influence each other. It is considered as an Autoregressive model because, each variable (Time Series) is modeled as a function of the past values, that is the predictors are nothing but the lags (time delayed value) of the series. The primary difference is those models are uni-directional, where the predictors influence the Y and not vice-versa. Whereas Vector Auto Regression (VAR) is bi-directional. That is, the variables influence each other.

Algorithm:

1) Analyze the time series characteristics.
2) Test for causation amongst the time series.
3) Test for stationarity.
4) Transform the series to make it stationary, if needed.
5) Find optimal order (p).
6) Prepare training and test datasets.
7) Train the model.
8) Roll back the transformations, if any.
9) Evaluate the model using test set.
10) Forecast to future.

D. VAR-IM:

VAR-IM is an improved VR algorithm for imputation of missing values using Expectation and Minimization(EM) algorithm and Prediction Error Minimization (PEM) method. VAR-IM provides improvements to speed and accuracy for imputing missing values of multivariate time series datasets. It outperforms the commonly used methods such as list wise deletion, linear regression imputation and EM algorithms. Since it uses the error minimization techniques along with VAR method it gives even better results than the VAR method.

E. Miss Forest:

MissForest is a machine learning-based data imputation algorithm that uses Random Forest algorithm. Stekhoven and Bulhmann, the creators of algorithm made a study in 2011 where various data imputations methods are used and MissForest outperformed them all the other algorithms by over 50%. First, the available values' mean/median/mode is calculated, and all the missing values are filled with the mean/median/mode values. The missing values are named Predict and available values are called training. The dataset is now without any missing values, it will now be fed to the random forest model. This process of looping through missing data points repeats several times, each iteration improving on better and better data. Iterations continue until some stopping criteria is met or after a certain number of iterations are done. Normally, datasets become well imputed after four to five iterations, but it depends on the size and amount of missing data.

CONCLUSION

GeoAI is an inter-disciplinary field which can be used in many fields which require geographical data. In order to get maximum efficiency, high quality data is a must. But most datasets which consist of big data have missing values. In order to solve this, we need to process data. Data cleaning can be done either using imputation or interpolation. We discussed about imputation methods. After the data processing, the dataset becomes more reliable, and the models could be trained more effectively. This leads to better forecasting and predictions.

Missing data is one of the most commonly found issue in any dataset and it is also considered as one of the most important issues that needs to be solved. While many of the algorithms can be able to perform with missing data, it will often lead to less accurate predictions and results.

For Geo AI algorithms to work and give accurate results, the data it needs to train on, needs to be perfect without flaws as well. But as expected, all of the data related to Earth's atmosphere, lithosphere, hydrosphere and biosphere are influenced by a plethora of variables. The data available with us is never accurate and has a lot of missing data.

In order to solve that problem, we have researched on missing data algorithms and found that the four algorithms found in this paper are some of the best algorithms to fight the missing data problem. And in them, MissForest algorithm is considered the best algorithm to use to fill the missing data in datasets.

But as technology continues to grow, there will be more breakthroughs and also more ways to dissolve the present issues will be found also with more streamlined algorithms.

REFERENCES

[1] National Academies of Sciences, Engineering, and Medicine, 2019. Leveraging artificial intelligence and machine learning to advance environmental health research and decisions: Proceedings of a workshop—In brief.

[2] Weichenthal, S., Hatzopoulou, M. and Brauer, M., 2019. A picture tells a thousand. . . exposures: opportunities and challenges of deep learning image analyses in exposure science and environmental epidemiology. Environment international, 122, pp.3-10.

[3] Gonzales-Inca, C., Calle, M., Croghan, D., Torabi Haghighi, A., Marttila, H., Silander, J. and Alho, P., 2022. Geospatial Artificial Intelligence (GeoAI) in the Integrated Hydrological and Fluvial Systems Modeling: Review of Current Applications and Trends. Water, 14(14), p.2211.

[4] VoPham, T., Hart, J.E., Laden, F. and Chiang, Y.Y., 2018. Emerging trends in geospatial artificial intelligence (geoAI): potential applications for environmental epidemiology. Environmental Health, 17(1), pp.1-6.

[5] Bhaskaran, K., Gasparrini, A., Hajat, S., Smeeth, L. and Armstrong, B., 2013. Time series regression studies in environmental epidemiology. International journal of epidemiology, 42(4), pp.1187-1195.
[6] Asha, P., Natrayan, L.B.T.J.R.R.G.S., Geetha, B.T., Beulah, J.R., Sumathy, R., Varalakshmi, G. and Neelakandan, S., 2022. IoT enabled environmental toxicology for air pollution monitoring using AI techniques. Environmental Research, 205, p.112574.
[7] Lokendra C. (2021) Geospatial AI/ML Applications and Policies: A Global Perspective. ISBN: 9789083156903
[8] Schmidt, C.W., 2020. Into the black box: what can machine learning offer environmental health research?.
[9] Temenos, A., Tzortzis, I.N., Kaselimi, M., Rallis, I., Doulamis, A. and Doulamis, N., 2022. Novel Insights in Spatial Epidemiology Utilizing Explainable AI (XAI) and Remote Sensing. Remote Sensing, 14(13), p.3074.
[10] 10. Pearce, N., Vandenbroucke, J. and Lawlor, D.A., 2019. Causal inference in environmental epidemiology: old and new. Epidemiology (Cambridge, Mass.), 30(3), p.311.
[11] Gregory, R.D. and van Strien, A., 2010. Wild bird indicators: using composite population trends of birds as measures of environmental health. Ornithological Science, 9(1), pp.3-22.
[12] [Kamel Boulos, M.N., Peng, G. and VoPham, T., 2019. An overview of GeoAI applications in health and healthcare. International journal of health geographics, 18(1), pp.1-9.
[13] Eaker, C., 2016. What Could Possibly Go Wrong? The Impact of Poor Data Management.
[14] McFarland, D.A. and McFarland, H.R., 2015. Big data and the danger of being precisely inaccurate. Big Data & Society, 2(2), p.2053951715602495.
[15] Starbuck, W.H., Hodgkinson, G.P. and Mezias, J.M., 2008. Decision making with inaccurate, unreliable data. The Oxford Handbook of Organizational Decision Making.
[16] Brown, A.W., Kaiser, K.A. and Allison, D.B., 2018. Issues with data and analyses: Errors, underlying themes, and potential solutions. Proceedings of the National Academy of Sciences, 115(11), pp.2563-2570.
[17] Azimi, S. and Pahl, C., Continuous Data and Model Quality Management for Machine Learning based Data-as-a-Service Architectures.
[18] Fu, R., Vandermeer, B.W., Shamliyan, T.A., O'Neil, M.E., Yazdi, F., Fox, S.H. and Morton, S.C., 2013. Handling continuous outcomes in quantitative synthesis.
[19] Tripepi, G., Jager, K.J., Stel, V.S., Dekker, F.W. and Zoccali, C., 2011. How to deal with continuous and dichotomic outcomes in epidemiological research: linear and logistic regression analyses. Nephron Clinical Practice, 118(4), pp.c399-c406.
[20] Joch, M., Dohring, F.R., Maurer, L.K. and M ¨ uller, H., 2019. In- ¨ ference statistical analysis of continuous data based on confidence bands—Traditional and new approaches. Behavior Research Methods, 51(3), pp.1244-1257.
[21] Bashir, F. and Wei, H.L., 2018. Handling missing data in multivariate time series using a vector autoregressive model-imputation (VAR-IM) algorithm. Neurocomputing, 276, pp.23-30