Deep Learning techniques on GIS Data to predict Social factors

Siddharth Satyam
SOC 479 Individual Project
Indian Institute of Technology, Kanpur
sidsa@iitk.ac.in

***Abstract*—The following research article is an individual project for course SOC479: Population, Economy And Society. The project is an interdisciplinary research oriented project that connects the sub-discipline of machine learning with demography and sociology, since, I have a strong background in the machine learning domain, although I hail from a mechanical engineering background.**

I. Introduction

With the advent of machine learning techniques, several disciplines have involved its use in data prediction, forecasting and generation. It has been highly used in healthcare, manufacturing, finance, business, and various other fields. Most machine learning models employed in these fields are supervised learning models and training them requires huge amounts of data. Data scientists involved in these fields spend a lot of time in collection and preparation of data. When we talk about data collection, the field where it is most systematically done on a large sample space is demography. The sources of demographic data are census, vital statistics data, administrative data, sample surveys, spatial data, and other forms of data collection methods. These enormous data sources can be easily used for data analysis and prediction of determinants in areas where the data collection has not been sufficient. Recently, there have been works involving linking COVID cases to socio-economic determinants using machine learning with the use of census data [3]. Electoral divisions were clustered according to the COVID cases spread among the population. In another work [4], an adult's income class was predicted based on certain set of attributes using UCI Adult dataset, a data collection measure in USA. In these works, emphasis has been to utilize the demographic data to determine measures that are not directly sampled, sampled but not properly available for each region, or are abstract entities. While not much work has been done on the Indian census data and other demographic data, our goal in this paper is to describe a methodology for utilizing a data collection measure to predict socio-demographic entities for regions where the sampling is not much informative as other nearby regions.

A. GIS (Geographical Information System)

GIS is a computer assisted system for the acquisition, storage, analysis and display of geographic data [2]. It stores spatially distributed data by geo-referencing the data i.e., the exact geographical coordinates are known where a data collection event occurred. It is widely used in agriculture, archaeology, environment, health, forestry, navigation, road and railways. The GIS associates data to *vector polygons* and geometry which represent the geographical features. As shown in Fig.1, a vector polygon is a single connected sequence of three or more co-planar lines which form a closed loop [1]. A line is defined by the vector coordinates of its two end points in $\mathbb{R}^2$. A building may be stored with the help of *Building Contour Polygons* and a neighborhood is defined by a *Neighborhood Geometry*. In [1], a deep learning technique has been described where they have discussed an elaborate methodology on utilizing the vector polygons in GIS. However, we will discuss on the applications of the model for prediction of socio-economic factors.

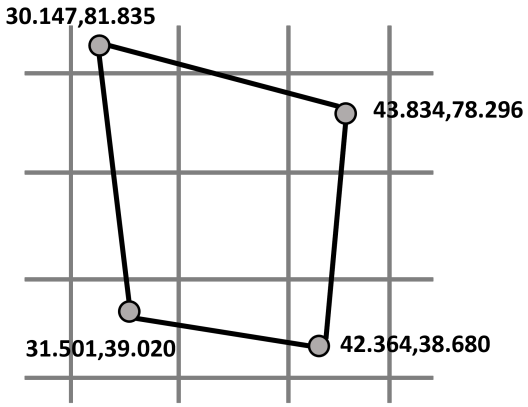

Fig. 1. A vector polygon on the coordinate system

B. Socio-Demographic dimension

Social issues such as domestic violence, female infanticide etc. have been prevalent in the country for long. With the newer generation getting educated, such issues have gradually decreased but are still existent in today's society. The rural areas tend to suffer greater than the urban areas and there are several factors that influence and determine these issues. The factors such as literacy rates, sex ratio, population density, ethnicity etc. generally reflect the prevailing issues in the society. Although the above stated factors are not directly the contributors and causal entities of such issues, they surely hold correlation.

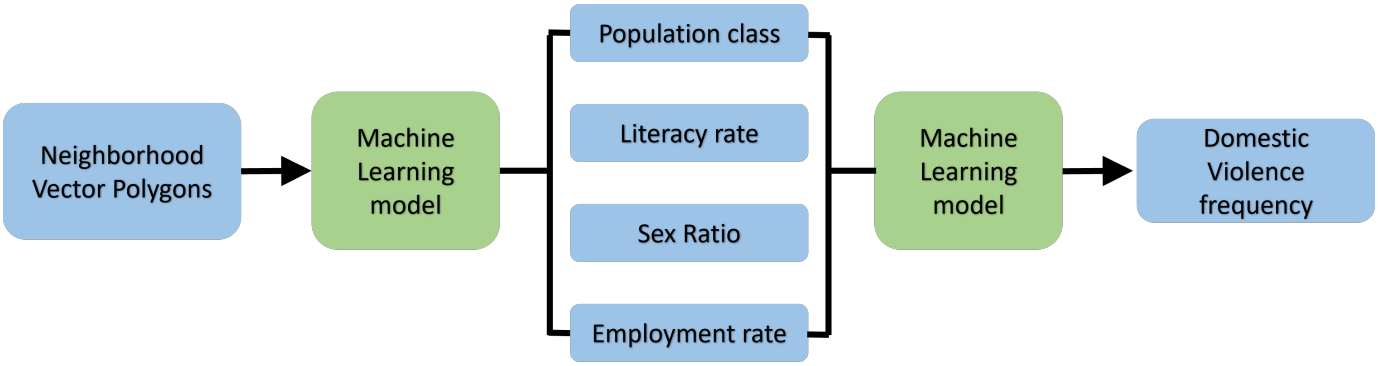

Fig. 2.

A greater literacy rate generally contributes to greater living standards although it might not be the sole reason for it, as there are other factors that play their roles. When we talk about correlation, a mathematical model can help us determine what role the factors play in domestic violence rates, for instance. The idea is based on the prediction of domestic violence rates based on its correlation to abstract entities such as vector geometry and its association with demographic data.

We can employ deep learning techniques on the vector geometries to generate a prediction for Domestic Violence frequency in specific Neighborhoods where adequate data has not been collected due to several reasons. There might have been possible cases where the victim did not report the incident due to fear or the victim was not permitted to report. Moreover, there have been many false cases reported in the recent times which could overstate the actual numbers. A predicted data could reflect the approximate seriousness of the issue in such areas and appropriate action could be taken thereby.

II. Methodology

As depicted in Fig. 2, the vector polygons (G^i) for the different neighborhoods will be the inputs. We will follow the normalization method in [1]. Each point vector (p_j^i) in a vector polygon (G^i) is first normalized as:

$$p_j^{i\prime} = \frac{p_j^i - \overline{p_j^i}}{s} \tag{1}$$

Where $\overline{p_j^i}$ is the geometry centroid of G^i and s is a scale factor. After normalization, a one-hot vector [1 0 0] is appended to each point vector to indicate a normal point, [0 0 1] to indicate the final stop point or [0 1 0] to indicate any sub-geometry. Finally, we will get a tensor as an input of size $(N \times 5)$ where N is the number of points in the polygon. We will then use a neural network with convolutional neural network(CNN) layers that will take the tensor $(N \times 5)$ corresponding to a geometry G^i as input. The output will be a two dimensional vector where each row will be a vector denoting feature attributes of each demographic entity, i.e.,population class (denoting the categories – high, low, medium), literacy rate, sex ratio and employment rates. Since the model training depends on the available data, we can add more demographic entities to the output field depending on the availability and its correlation with domestic violence frequency. The first neural network will have CNN layers, with the first layer having a kernel size of k, filter size of f and stride of 1, so that it produces an output of f dimensional vectors of demographic entities. The kernel size of k would denote a sliding window over k geometries at once, striding one at a time. Thereafter, the two dimensional output vector will be fed to a max pooling layer which will produce an output that will be used to produce the domestic violence frequency.

III. Conclusion

The idea if implemented, will be unique in terms of its application to socio-economic entities. There are many rural areas where data collection is not easy and so this idea can provide a gateway to utilize the already existing data trends of highly surveyed areas to give an approximately correct picture of the scenario in the rural areas and areas with difficult transportation. The merits also include no additional infrastructure cost as it utilizes the existing GIS hardware. In addition to socio-economic entities, information such as water resource management could also be utilized, as plans have been proposed to collect water resource data in rural areas as in [2]. The demerits might include the tedious process of associating the data with spatial coordinates in the GIS system, but since it is being implemented in the country for a while, we can say that our idea is feasible.

References

[1] R.H. van 't Veer et.al.,"Deep Learning for Classification Tasks on Geospatial Vector Polygons",arXiv:1806.03857v2 [stat.ML] 11 Jun 2019

[2] Nayak, S.K., Thorat, S.B., Kalyankar, N. "GIS: Geographic Information System An application for socio-economical data collection for rural area." ArXiv, abs/1004.1793., (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, No. 3, March 2009

[3] Ghahramani, Mohammadhossein Pilla, Francesco. (2021). "Leveraging Artificial Intelligence to Analyze the COVID-19 Distribution Pattern based on Socio-economic Determinants."

[4] N. Chakrabarty and S. Biswas, "A Statistical Approach to Adult Census Income Level Prediction," 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2018, pp. 207-212, doi: 10.1109/ICACCCN.2018.8748528.