**Universidade de Brasília**
**Departamento de Estatística**

**Research Project (Bachelor Dissertation)**
**PE-GNN**

**Rafael de Acypreste**

Projected submitted to the Department of Statistics at the University of Brasília, as part of the requirements to obtain the Bachelor's Degree in Statistics.

**Brasília**
**2024**

**Rafael de Acypreste**

# Research Project (Bachelor Dissertation)
# PE-GNN

Orientador(a): Prof. Professora

Projected submitted to the Department of Statistics at the University of Brasília, as part of the requirements to obtain the Bachelor's Degree in Statistics.

**Brasília**
**2024**

# Contents

# 1 Introduction

Many sources of statistical information use geographical or, at least, potential geographical information. In some problems, considering a global measure under a homogeneous assumption is enough to reach reliable inferences. However, when variability can be traced by location, it is necessary to consider the spatial dependence of the data. In this case, global estimations can be biased and inefficient in inferring about local conditions[1]. The evolution of diseases, housing price variation, or climate change are examples of problems that can be more accurately understood by considering spatial dependence. In those situations, multi-valued statistics are more suitable to represent the data.

The main topic of this dissertation is tackling the problem of modeling geospatial data, represented by a graph. When tackling graph information, the first element to define is a graph. Mathematically, for every use of graphs, it is necessary to define accurately its structure, given by “nodes” ($N$), “edges” ($E$)[2], and a graph ($G(N, E)$), properly speaking. In the real world, some topics can be naturally suited to graphs further than geospatial data. For example, social networks, where people are nodes and the relations between them are the edges. Given a protein environment, the nodes are representations of the amino acids and edges of the interaction between them. In observational studies, causal relations can be modeled by a “directed acyclic graph” (Pearl, 2009), whose variables are nodes and causal-effect directions are represented by the edges.

The traditional statistical approaches are unable to tackle spatial autocorrelation. It is very common for spatial variables to exhibit spatial non-stationarity (Brunsdon; Fotheringham; Charlton, 1996). Then, the most popular statistical models might give non-random residuals leading to wrong conclusions. In this context, Graph Neural Networks (GNN) have been applied to deal with this complex geographic information, where data used to be highly correlated (Klemmer; Safir; Neill, 2023, p. 2). The main feature of a GNN is to learn a representation of a set of graph-structured pieces of information. Moreover, the intuition in this learning process is to update the node representation based on its features and neighbor representations (Tang; Liao, 2022, pp. 41–3).

GNNs are a class of neural networks that operate on graphs and are also known as deep learning on graphs, graph representation learning, or geometric representation learning (Wu *et al.*, 2022a, p. vii). They are a widely adopted tool to represent connected structures and relations. This class of model deals with three graph data singularities: 1) variable topology; 2) variable size (network graphs may have billions of points, for

[1] Some authors (Brunsdon; Fotheringham; Charlton, 1996) distinct local and spatial pieces of information. Local estimation aims to treat commonly non-linear relations that differ along the feature space independently of geographical information. LOWESS regression is a method to tackle local variation.

[2] The information stored in nodes and edges is called node and edge embedings, respectively (Prince, 2024, p. 243).

example); 3) often, there graph is monolithic; thus, testing in new data is not always feasible (Prince, 2024, p. 240).

Therefore, large graphs with millions of nodes became intractable in the traditional frameworks, especially due to high computational demand in process nodes and edges relations, low parallelizability, given the multiplicity of interconnections between nodes, and inapplicability of deep learning methods like using samples of data in the learning process that could lose nodes interconnections (Cui *et al.*, 2022, p. 18). Furthermore, graphs do not present regular structures like images or texts do. Consequently, well-defined neural network architectures such as convolutional or recurrent neural networks are not well suited to deal with graphs (Cui *et al.*, 2022; Wu *et al.*, 2022b).

A GNN is a neural network that can represent geographical information and spatial dynamics. It can represent information like points of interest, traffic speed at a specific location, etc. (Klemmer; Safir; Neill, 2023). The first task is to find an adequate and accurate graph representation in low dimensions (Cui *et al.*, 2022, p. 17). These representations should permit the identification of patterns, structures, and data generation processes for analysis and prediction.

However, GNNs might be insufficient for modeling complex spatial effects. In a graph, there are two proximities of interest (Cui *et al.*, 2022, p. 21): first-order proximity indicates the observed distance between two nodes; second-order proximity refers to the "context" (neighbors) of the two nodes. Then, some task is to find a method to make graph embedding to tackle both proximities in a low dimensional space. In this context, Klemmer; Safir; Neill (2023) proposed the "Positional Encoder Graph Neural Network" (PE-GNN).

A traditional couple of geographical references (latitude/longitude) usually is not the feature to input a GNN. Setting some k-neighborhood became a hyperparameter subject to exhausting tests. Therefore, discovering any latent variable — probably a high-dimensional one — from geographical coordinates might be more suitable for training a GNN (Klemmer; Safir; Neill, 2023, p. 3). PE-GNN consists of using a positional encoder for contextual embedding for a point in space, which is concatenated with other features in the GNN training process. Doing this, the prediction became more accurate, supported by the calculated spatial autocorrelation. Then, the graph information is presented jointly with the learned positional embeddings. It works with any GNN singular structure (Klemmer; Safir; Neill, 2023, pp. 1–2).

This ability to transform dimensions of the geographical features constitutes a field of the so-called "representation learning", in which the main objective is "[...] to extract sufficient but minimal information from data" (Zhao *et al.*, 2022, p. 3). When done by a human, is also called "feature engineering". However, it is a high-cost process and expert-dependent. Then, one goal of deep learning — and positional encoder in this case — is to make representation learning more abstract, useful (Zhao *et al.*, 2022, p. 4), and less human-dependent.

A Convolutional Graph Neural Network (*GCN* for short) is convolutional in the sense that each layer updates its parameters aggregating the node context (neighborhood) (Prince, 2024, p. 248).

# 2 Further study points

ST (C, $_{min}$ , $_{max}$) is a sinusoidal transformation. What are these $\sigma$ about in Klemmer; Safir; Neill (2023)?

# 3 Objetives

## 3.1 General objective

Refactor PE-GNN model to include local clusters information.

## 3.2 Specific objectives

-

# 4 Literature Review

# 5 Methods

# 6 Schedule

The activities to be developed during 1/2024 are:

Table 1: Timeline

| Activities | 1/2024 | | | | |
|---|---|---|---|---|---|
| | Mar | Apr | May | Jun | Jul |
| Linear Algebra revision | ■ | ■ | ■ | ■ | |
| Literature review about GNN architectures | ■ | ■ | ■ | | |
| Python advanced programming courses | | ■ | ■ | ■ | ■ |
| Analysis of the startpoint dissertation | | | ■ | ■ | ■ |
| Replication of the original paper and dissertation codes | | | | ■ | ■ |
| Writing of the research proposal | | | ■ | ■ | ■ |
| Incremental fine-tunings planning | | | | | ■ |
| Drawing um of the partial report | | | | | ■ |

In the 2/2024 season, the activities are:

Table 2: Timeline

| Activities | 2/2024 | | | | |
|---|---|---|---|---|---|
| | Aug | Sep | Oct | Nov | Dec |
| Adaptation of code to change functions and parameters and documentation | ■ | ■ | ■ | ■ | |
| Training and test of the model with distinct databases | | | ■ | ■ | |
| Drawing up of the new algorithms and mathematical models | | | | | |
| Model fine-tuning | | | ■ | ■ | |
| Results and discussion sections writing | | | ■ | ■ | ■ |
| Model publication in the GitHub platform | ■ | ■ | ■ | ■ | ■ |
| Monograph defense | | | | | ■ |