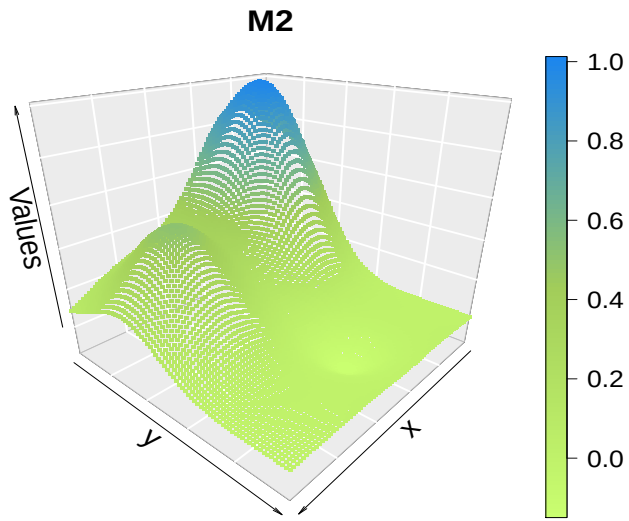
Statistical methods for spatial data analysis

Lecture 11: Introduction to Machine Learning for spatial data

Joaquin Cavieres

Geoinformatics, Bayreuth University



Outline

1. Introduction

Spatial data is any type of data that directly or indirectly references a specific geographical area or location.

Spatial data also are called "geospatial" data and the advantage of this type of data is that they can numerically represent a physical object in a geographic coordinate system.

Some examples of spatial data:

- GIS data: maps, satellite imagery, aerial photographs, elevation models, and land cover/land use data.
- GPS data: It is commonly used for navigation, tracking, and mapping purposes.
- Remote sensing: It capture data about the Earth's surface from a distance, typically using satellites or aircraft.
- Sensor Networks: it is placed in a physical environment can collect spatial data such as temperature, humidity and air quality.
- Environmental data: Spatial data related to environmental phenomena, such as climate data, weather patterns, pollution levels, and ecological habitat maps.

What is the importance of spatial data?

What is the importance of spatial data?

Urban planning and development: It helps identify suitable locations for infrastructure projects, such as roads, buildings, and parks, by considering factors like land use, population density, transportation networks, and environmental constraints.

Figure 1: Source: National urban development `https://www.bmi.bund.de/EN/topics/building-housing/city-housing/national-urban-development/national-urban-development-node.html`

Environmental management: It helps track changes in land cover, analyze the impact of deforestation or urbanization, identify areas prone to natural disasters, and plan conservation efforts.

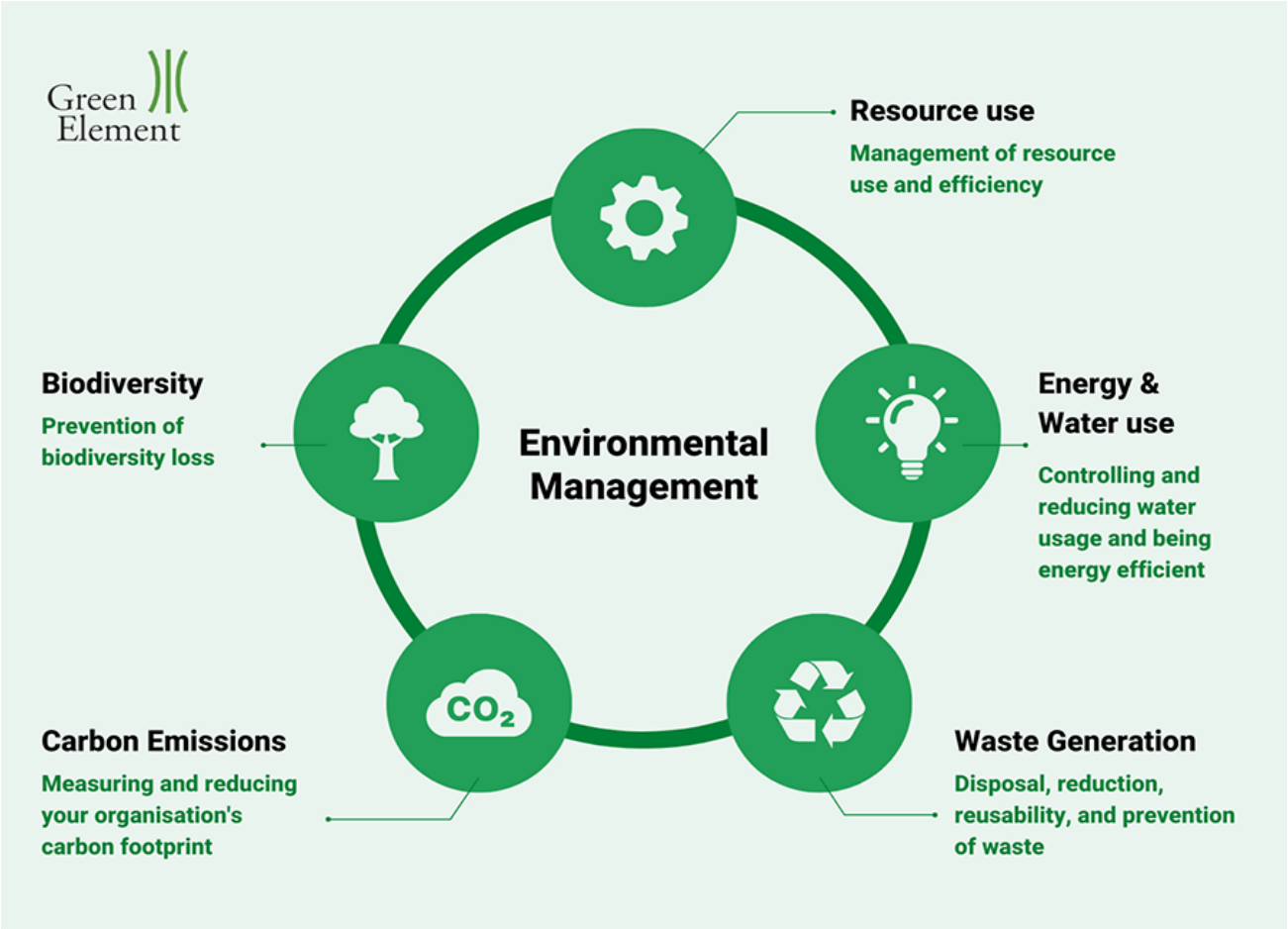


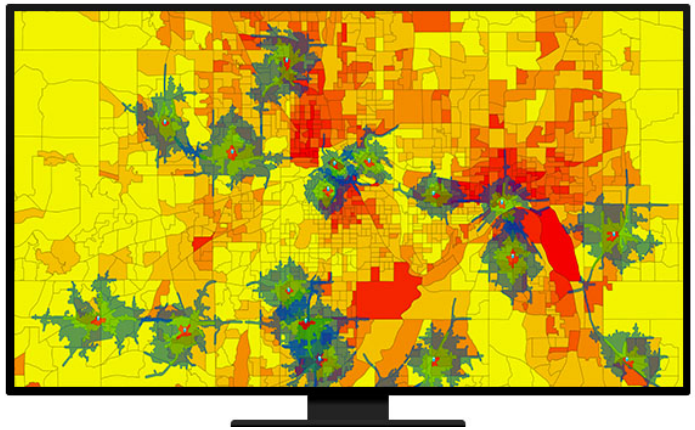
Figure 2: Elements considered when implementing an Environmental Management System. Source: `https://www.greenelement.co.uk/blog/what-is-an-environmental-management-system/`

Transportation and logistics: It enables the optimization of transportation networks, including road networks, public transportation routes, and logistics supply chains.

Figure 3: Transport logistic. Source: `https://transportlogistic.de/en/trade-fair/industry-insights/`

Market analysis and retail planning: It helps identify suitable locations for retail stores based on factors such as population demographics, competitor locations, accessibility, and consumer behavior.

(a)



(b)

Figure 4: Understand where your customers come from. Source: `https://www.esri.com/en-us/industries/retail/strategies/market-research-planning`

2. Introduction to Machine Learning (ML)

A brief definition:

ML (or statistical learning) is concerned with the use of statistical and computational models for identifying patterns in data and predicting from these patterns.

2.1. Train data and model

Train data

Refers to the set of examples or observations that is used to teach a ML algorithm. It consists of input data and corresponding output labels or target values.

Key points about train data:

- **Input Data**: Also known as features or predictors, are the variables or attributes that are provided to the ML algorithm. These can be numerical values, categorical variables, text, images, or any other relevant data type.
- **Response variable**: It represents the desired outcome or prediction that the ML algorithm should learn to generate based on the input data. The labels can be categorical (classification task) or continuous (regression task).

Model

A model, in the context of ML (statistical learning), is a representation of the learned patterns and relationships extracted from the training data.

Key points about the model:

- **The model**: It refers to the specific algorithm or technique used to represent and learn from the training data.
- **Evaluation and validation**: Once the model is trained, it is evaluated using separate validation or test data to assess its performance. This evaluation helps determine the model's effectiveness, identify potential issues like overfitting or underfitting, and make necessary adjustments.

2.2. Supervised and unsupervised learning

Supervised learning

It is a type of a ML approach in which the algorithm learns from labeled training data to make predictions or decisions.

The algorithm learns the relationship between the input data and the output labels by minimizing the error or discrepancy between its predicted outputs and the true labels.

Supervised learning is commonly used for tasks such as classification, regression, and sequence labeling, where the algorithm learns to map input data to specific target labels or values based on the labeled training data.

Unsupervised Learning

It is a type of ML approach where the algorithm learns patterns and structures in unlabeled data without any specific output labels or target values.

In unsupervised learning, the algorithm explores the inherent structure and relationships within the data to discover meaningful patterns, clusters, or representations.

Unsupervised learning is commonly used for tasks such as clustering, anomaly detection, data visualization, and feature learning, where the focus is on exploring and understanding the data intrinsic structure without relying on explicit labels or target values.

3. Application of ML to spatial data

ML is valuable for spatial data analysis due to several reasons:

(I) Complex patterns and relationships.
(II) Scalability and efficiency.
(III) Automated feature extraction.
(IV) Spatial data fusion
(V) Prediction and forecasting.

3.1. Examples of ML applied to spatial data

Classification

- Land cover classification.
- Species habitat classification.

Regression

- Price prediction.
- Air quality prediction.

Clustering

- Spatial segmentation.
- Customer Segmentation.

4. Common ML algorithms for spatial data

Some ML algorithms for spatial data:

- Decision trees
- Random forests
- Support Vector Machines (SVM)
- Neural networks
- k-Nearest Neighbors (kNN)
- Gaussian processes in 2D (aka, Gaussian random field)
- Spatial regression models

5. Challenges and considerations

1. Spatial autocorrelation

Spatial autocorrelation violates the assumption of independence made by many ML algorithms, leading to potential biases or misleading results.

1. Spatial autocorrelation

Spatial autocorrelation violates the assumption of independence made by many ML algorithms, leading to potential biases or misleading results.

Solution?

1. Spatial autocorrelation

Spatial autocorrelation violates the assumption of independence made by many ML algorithms, leading to potential biases or misleading results.

Solution?

1. Spatial autocorrelation

Spatial autocorrelation violates the assumption of independence made by many ML algorithms, leading to potential biases or misleading results.

Solution?

Spatial regression models or spatially aware algorithms.

2. Spatial heterogeneity

Spatial data in different regions or subregions have distinct characteristics or behaviors (heterogeneity).

2. Spatial heterogeneity

Spatial data in different regions or subregions have distinct characteristics or behaviors (heterogeneity).

Solution?

2. Spatial heterogeneity

Spatial data in different regions or subregions have distinct characteristics or behaviors (heterogeneity).

Solution?

2. Spatial heterogeneity

Spatial data in different regions or subregions have distinct characteristics or behaviors (heterogeneity).

Solution?

Geographically Weighted Regression (GWR) or Geographically Weighted Random Forest (GRF)

3. Scale and resolution

Spatial data can exist at different scales and resolutions, ranging from global to local levels. Matching the scale and resolution of the data with the appropriate modeling techniques is crucial to ensure meaningful analysis.

3. Scale and resolution

Spatial data can exist at different scales and resolutions, ranging from global to local levels. Matching the scale and resolution of the data with the appropriate modeling techniques is crucial to ensure meaningful analysis.

Solution?

3. Scale and resolution

Spatial data can exist at different scales and resolutions, ranging from global to local levels. Matching the scale and resolution of the data with the appropriate modeling techniques is crucial to ensure meaningful analysis.

Solution?

3. Scale and resolution

Spatial data can exist at different scales and resolutions, ranging from global to local levels. Matching the scale and resolution of the data with the appropriate modeling techniques is crucial to ensure meaningful analysis.

Solution?

Upscaling or downscaling of data may be required.

4. Computational requirements

Spatial data can be vast and require substantial computational resources for processing and analysis.

4. Computational requirements

Spatial data can be vast and require substantial computational resources for processing and analysis.

Solution?

4. Computational requirements

Spatial data can be vast and require substantial computational resources for processing and analysis.

Solution?

4. Computational requirements

Spatial data can be vast and require substantial computational resources for processing and analysis.

Solution?

Efficient algorithms, distributed computing frameworks, or parallel processing techniques may be needed.

5. Interpretability and explainability

ML models are considered black box models, making it challenging to interpret or explain the reasoning behind their predictions or decisions.

5. Interpretability and explainability

ML models are considered black box models, making it challenging to interpret or explain the reasoning behind their predictions or decisions.

Solution?

5. Interpretability and explainability

ML models are considered black box models, making it challenging to interpret or explain the reasoning behind their predictions or decisions.

Solution?

5. Interpretability and explainability

ML models are considered black box models, making it challenging to interpret or explain the reasoning behind their predictions or decisions.

Solution?

??

6. Conclusion

Advantages:

- ML allows us to discover complex patterns, handle large-scale data, and automate feature extraction.
- More common ML for spatial data are: classification, regression and clustering.

Disadvantages:

- Considers the spatial autocorrelation, heterogeneity, data quality issues, computational requirements and interpretability of the modelling process.

Thanks!...


Williams, J.M., 2022. Spatial Machine Learning: An intro in applying machine learning techniques to spatial data with R. Source: https://justinmorganwilliams.medium.com/spatial-machine-learning-29137dcd1f5f


R companion to Geographic Information Analysis. Spatial Data Science with Applications in R and terra. Source: https://rspatial.org/


Pebesma, E and Bivand, R., 2023. Spatial Data Science with applications in R. Source: https://rspatial.org/