# Neural Networks for Spatial Models

## MASHS 2026 Conference

Cécile Hardouin[1] Jean-Charles Lamirel[2] Castillo Sossou Akpa

[1]Modal'x, University Paris Nanterre, Nanterre, France
hardouinc@parisnanterre.fr

[2]Synalp, LORIA, Vandoeuvre-lès-Nancy, France
lamirel@loria.fr

September 11, 2026

# Outline

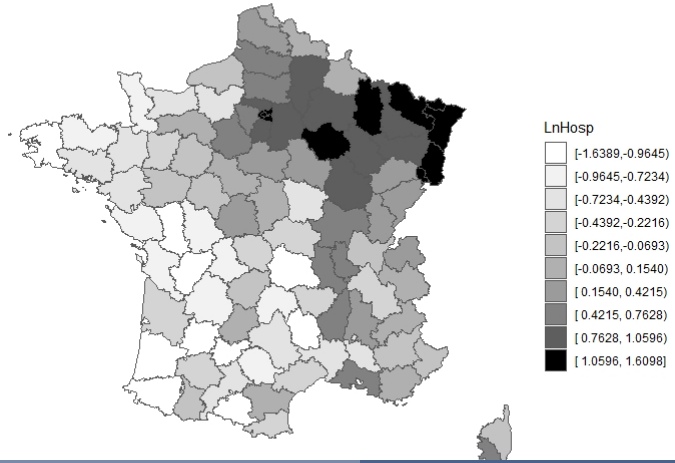

1. Introduction & Problem Statement
2. Spatial Econometric Models
3. Neural Clustering Algorithms
4. Applications & Experimental Results
5. Conclusion & Future Work

# Introduction & Spatial Context

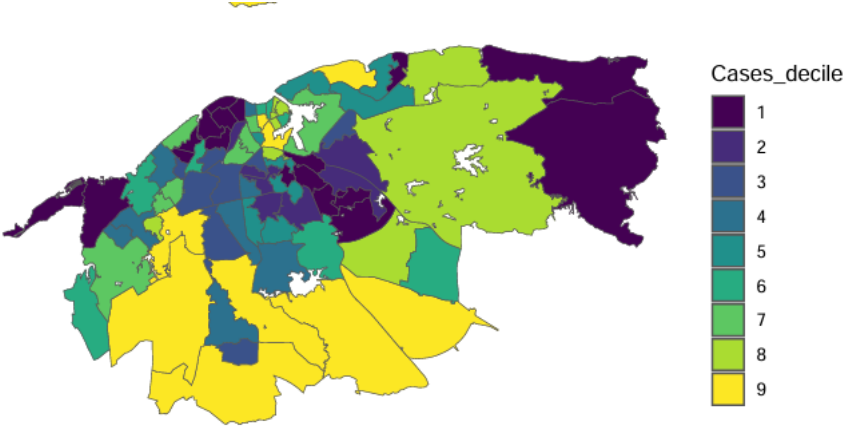
- **Ubiquity of Spatial Data:**
  - Epidemiology, economics, sociology, image analysis, environmental sciences.
  - Frequent presence of **spatial dependence** (spatial autocorrelation).
- **Tobler's First Law of Geography:**
  *"Everything is related to everything else, but near things are more related than distant things."*
- **Limitations of Classic Regression Models:**
  - Independence of observations (standard OLS assumption) is violated.
  - GLS estimation (need a good estimation of $\Sigma_\varepsilon$)
  - For data on a lattice, we incorporate the spatial structure via a **neighbourhood graph.**

# Example Covid France

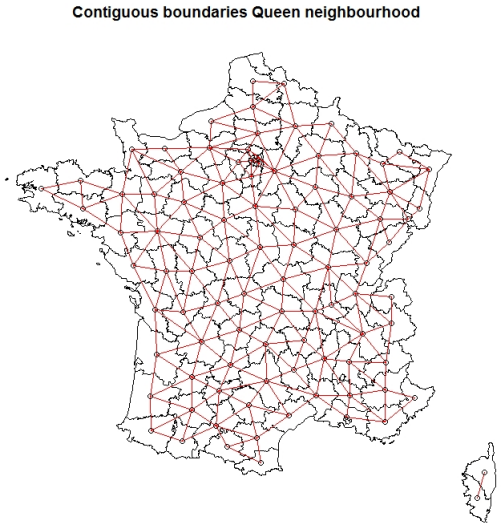
Log Hospitalization rate for Covid-19 by department (by decile)

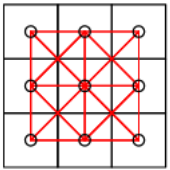


# Example Covid Habana

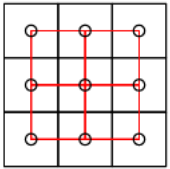


# Neighbourhood graphs



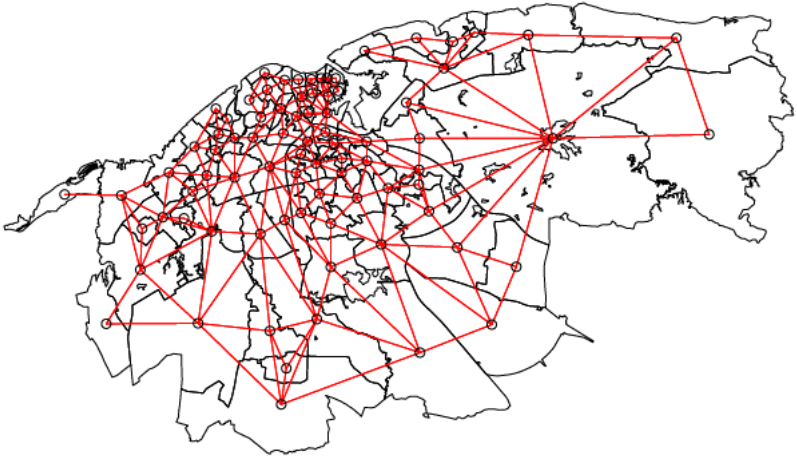



(a) Queen



(b) Rook

# Neighbourhood graphs



# Simultaneous Autoregressive Models (SAR, SEM, SAC, SDM, SDEM)

The neighbourhood graph is included in the models via a neighbourhhod matrix $W$.

## 1. Mixed Spatial Autoregressive Model (SAR / Spatial Lag)

$$Y = \rho WY + X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I)$$

- Directly models spatial effects in $Y$.
- Captures direct spatial interactions and feedback loops across the entire spatial network via the multiplier $(I_n - \rho W)^{-1}$.
- Can yield biased estimates if spatially correlated exogenous variables or omitted factors are left out of the specification;
- Coefficients $\beta$ cannot be interpreted as simple marginal effects due to spatial multiplier dynamics.

# Simultaneous Autoregressive Models (SAR, SEM, SAC, SDM, SDEM)

## 2. Spatial Error Model (SEM)

$$Y = X\beta + u, \quad u = \lambda Wu + \varepsilon$$

- Captures spatial autocorrelation in unobserved factors (error term).
- Coefficients $\beta$ represent direct marginal effects without spatial feedback loops.
- No Exogenous Spillovers

## 3. General Spatial Model (SAC / SARAR)

$$Y = \rho W_1 Y + X\beta + u, \quad u = \lambda W_2 u + \varepsilon$$

Combines spatial lag and spatial error, but Requires two distinct weight matrices $W_1 \neq W_2$.

- Simultaneously addresses endogenous spatial spillovers ($\rho$) and unobserved spatial error structure ($\lambda$).
- Nests both the SAR model (when $\lambda = 0$) and the SEM model.
- Identification Challenges, requires different weight matrices
- Computationally intensive and convergence issues.

# Simultaneous Autoregressive Models (SAR, SEM, SAC, SDM, SDEM)

There are other models like the Spatial Durbin model (SDM)or Spatial Durbin Error model (SDEM):

$$Y = \rho WY + X\beta + WX\theta + \eta$$

They are used whenever we want to model spilllovers from exogeneous variables, and they mitigate omitted variables bias.

But they frequently have multicolinearity issues.

# Spatial autoregressive models

**All these models are sensible to Matrix Choice** $W$:

Misspecification of the spatial weight matrix $W$ leads to biased error parameter estimates and distorted coefficient estimates.

Usually, $w_{ij}$ depends purely on geographical distance ($w_{ij} = d_{ij}^{-\alpha}$) or contiguity (Queen/Rook).

It is often row standardized: Observation $j$ is a weighted average of its neighbours.

# Problem Statement & Objectives

## The Challenge of Spatial Weight Matrices $W = (w_{ij})$

**Key Questions:**

1. What if two distant regions are socio-demographically or economically very similar?
2. How to choose a second weight matrix $W_2$ in general spatial models (SAC) without collinearity issues?
3. How can spatial econometrics be applied to non-geographically referenced data?

## Our Proposal

Use distances in the covariate space resulting from **neural clustering** (SOM, NG, GNG) to build neighbourhood matrices $W$.

# Recall 1: Kohonen Self-Organizing Maps (SOM)

- **Fixed Topological Grid:** 2D grid topology (e.g., $4 \times 3$ hexagonal).
- **Two-Step Learning Procedure:**
  1. *Competition:* Select the Best Matching Unit (BMU) $c(x)$:
  $$c(x) = \arg\min_i \|x - w_i\|$$
  2. *Cooperation & Adaptation:* Update the BMU and its topological neighbours on the grid:
  $$w_i(t+1) = w_i(t) + \alpha(t) h_{ci}(t)[x(t) - w_i(t)]$$
- **Key Property:** Topology preservation (similar inputs map to the same or adjacent neurons).

**SOM Neural Distances**

The distances $d_{K,ij}$ between observations $i$ and $j$ are computed as the distances between their corresponding prototype vectors on the map.

# Recall 2: Neural Gas (NG)

- **Motivation (Martinetz & Schulten, 1991):** Replace the fixed topological lattice of SOM with dynamic adaptation based purely on distance in the input space.
- **Neighbourhood Ranking Process:** For each input $x$, neurons are ordered by their proximity to $x$: $k_i(x, w)$.
- **Adaptation Rule:**

$$w_i(t+1) = w_i(t) + \epsilon(t) \cdot e^{-k_i/\lambda(t)} \cdot (x(t) - w_i(t))$$

- **Advantage vs. SOM:** Better representation of complex/non-linear data distributions.
- **Drawback:** The number of neurons $K$ remains a fixed pre-defined hyperparameter.

# Recall 3: Growing Neural Gas (GNG)

- **Evolving Network Topology (Fritzke, 1995):** Dynamically adapts the number of neurons and connections during learning.
- **Dynamic Mechanisms:**
  - *Insertion:* Adds a new neuron every $T$ iterations between the two neurons with the largest accumulated error.
  - *Deletion:* Connections have an "age". Connections exceeding maximum maturity are removed, as are isolated neurons.
- **Node Adjustment:** Strong adaptation for the winning unit and weaker adaptation for connected neighbours.
- **Optimal Model Detection** (Dugué, Lamirel & Chen, 2021): Automatically identifies optimal cluster count.

# Construction of Neural Spatial Weights $W_N$

- **Base Neighbourhood Graph:** Queen contiguity graph.
- **Weight Elements Specification:** For neighbouring sites $i$ and $j$ in the graph:

$$w_{ij} = d_{N,ij}^{-1} \quad \text{or} \quad w_{ij} = \frac{1}{\sqrt{d_{N,ij}}}$$

  where $d_{N,ij}$ is the Euclidean distance between prototypes corresponding to $i$ and $j$ after Neural clustering on covariates $X_1, \ldots, X_p$.
- **Normalization:** Each matrix $W$ is normalized by its largest eigenvalue ($\lambda_{\max}$), ensuring comparability and process stationarity.

# Our study

- We compare $W_N$ and $W_G$ in different applications, with $W_N = W_K, \quad W_N = W_{GNG}, \quad W_= W_N \times W_G$.
- SAC Model Specification:
  - $W_1 = W_G$ (Geographical distance-based matrix)
  - $W_2 = W_K$ or $W_{GNG}$ (Neural distance-based matrix)
- We apply spatial autoregressive models to non spatial data, with $W = W_K$ or $W_{GNG}$

# Application 1: COVID-19 Epidemic in French Departments

**Dataset Description:**

- **96 Departments** in Metropolitan France.
- **Dependent Variables ($Y$):**
  - Hospitalization rate (log)
  - Hospital death rate (log)
  - Death excess rate
- **Socio-economic Covariates ($X$):** Population density, ratio of top 3 cities, second homes rate, inactive rate, workers rate, emergency services, physicians.

# Application 1: COVID-19 Epidemic in French Departments

We compare matrices $W$:

$$W_{ij} = 1/\sqrt{d_{ij}^G} \quad \text{if } i \sim j \text{ (Queen)}$$

$$W_{ij} = 1/\sqrt{d_{ij}^K} \quad \text{if } i \sim j \text{ (Queen)}$$

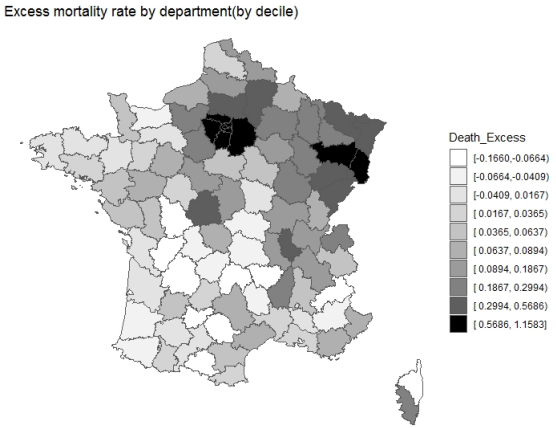
$$W_{ij} = 1/\sqrt{d_{ij}^{GNG}} \quad \text{if } i \sim j \text{ (Queen)}$$

We check that all N-distances between codes are $> 1$; then $W_{ij}^N = 1$ if $i$ and $j$ belong to the same neuron, and $W_{ij}^N = 1/\sqrt{d_{sl}^N}$ if $i$ belongs to neuron $s$ and $j$ to neuron $l$.

For each dependent variable $Y$, each model SAR, SEM, SAC, is estimated with the different matrices.

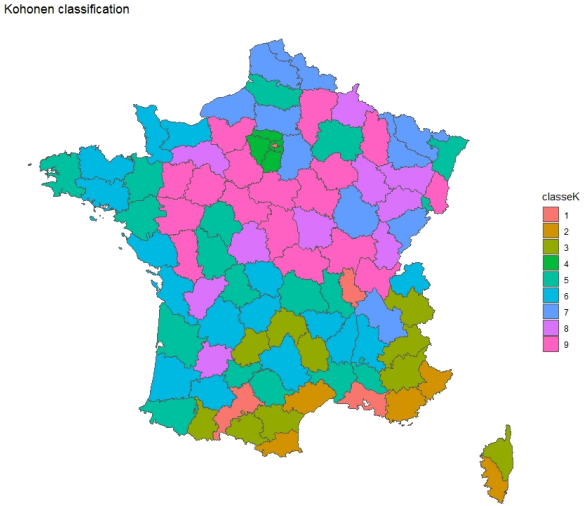
# Application 1: COVID-19 Epidemic in French Departments

Deciles distribution of `Death Excess`



# Clustering algorithms

Kohonen SOM: an hexagonal map of size 3 by 3



# Results: SAR Models for Death Excess

| **Variable** | **SAR** $W_{1/\sqrt{d_G}}$ | **SAR** $W_{1/\sqrt{d_K}}$ **(Kohonen)** | **SAR** $W_{1/\sqrt{d_{GNG}}}$ **(GNG)** |
|---|---|---|---|
| (Intercept) | -0.544 (0.030) | -0.559 (0.020) | -0.535 (0.027) |
| Ln Pop Density | 0.052 (0.024) | 0.077 (0.00016) | 0.086 ($2 \cdot 10^{-5}$) |
| Second Home | 0.001 (0.600) | 0.002 (0.293) | 0.002 (0.253) |
| Inactive | 1.044 (0.259) | 0.490 (0.576) | 0.346 (0.695) |
| Work | 2.966 (0.0048) | 2.494 (0.0117) | 2.235 (0.0246) |
| Emergency | -0.002 (0.561) | -0.002 (0.624) | -0.001 (0.729) |
| Physician | -0.087 (0.460) | -0.090 (0.425) | -0.118 (0.295) |
| $\rho$ **(Autocorrelation)** | **0.8092** ($5 \cdot 10^{-11}$) | **0.7326** ($2 \cdot 10^{-12}$) | **0.7097** ($4 \cdot 10^{-11}$) |
| **AIC** | **-80.581** | **-86.948** | **-85.499** |

Table: Regression coefficient estimates ($p$-value in parentheses).

# France COVID-19 Findings Analysis

- **Covariate Selection:** Consistently selects the same significant covariates across models: `Population Density` $(+)$ and `Workers` Ratio $(+)$.
  - For $Y =$*hospitalization*, the best model is a SAR model, with $W_K$
  - For $Y =$*death*, the best model is a SEM model, with $W_K$
  - For $Y =$*Death Excess*, the best model is a SAR model, with $W_K$
- We note that the coefficients of *emergency service* and *physician* variables are negative.
- **Performance Gain:** The **AIC** criterion constantly improves from the use of geographic distances to the use of neural distances.
- **General Spatial Model (SAC):** Simultaneous modeling with $W_1 = W_G$ and $W_2 = W_K$ confirms SAR model adequacy for hospitalizations/death excess, and SEM model for hospital deaths.

# Application 2: the Baltimore data

| | |
|---|---|
| PRICE | Sales price of house in 1,000 USD |
| NBATH | Number of bathrooms |
| NSTOR | Number of stores |
| GAR | Number of car spaces in garage (0 = no garage) |
| AGE | Age of dwelling in years |
| Ln(LOTSZ) | Lot size in hundreds of square feet |
| Ln(SQFT) | Interior living space in hundreds of square feet |
| PATIO | 1 if patio, 0 otherwise |
| FIREPL | 1 if fireplace, 0 otherwise |
| AC | 1 if air conditioning, 0 otherwise |

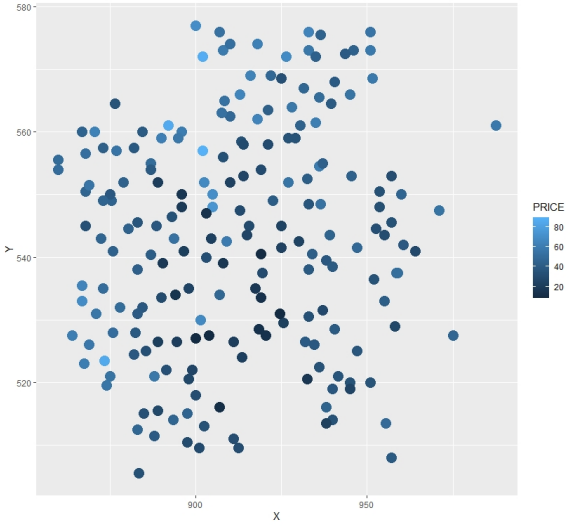
# Application 2: the Baltimore data

**Extension for categorical variables:**
In the usual K algorithm, each variable is centered and scaled.

- Each categorical variable with $k$ values is transformed into $k$ dummies
- The $k$th dummy is centered and scaled by $1/\sqrt{\frac{n}{n_k}}$ (Chavent et al. 2017)

# Application 2: the Baltimore data

**Sales price in 1,000 USD**



# Application 2: the Baltimore data

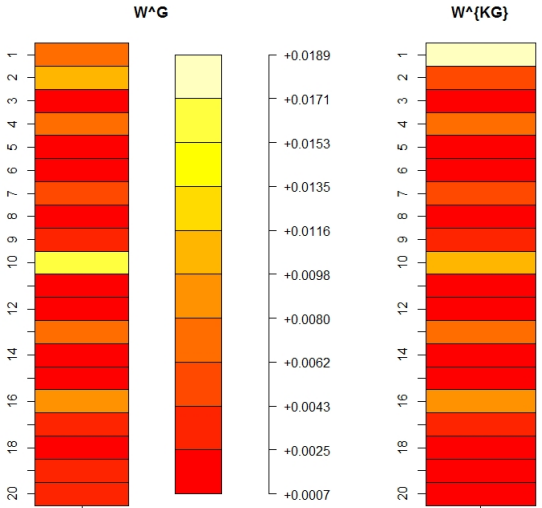
**Our design for $W$:**

- Geographic distances: $w_{ij}^G = 1/d_{ij}^G$
- Kohonen distances: $W_{ij}^K = 1$ or $1/d_{sl}^K$ if $i$ belongs to neuron $s$ and $j$ to neuron $l$. (hexaginal grid $(4 \times 4)$
- $W_{ij}^{GK} = 1/d_{ij}^G \times 1/d_{ij}^K$

Matrices are normalized by their largest eigen value.

# Application 2: the Baltimore data

**Weight matrices**



# Application 2: the Baltimore data

| Model | SEM $W_{1/d^G}$ | SEM $W_{1/d^{GK}}$ |
|---|---|---|
| Intercept | -4.948 (0.497) | -0.232 (0.977) |
| NBATH | 2.862 (0.042) | 2.779 (0.050) |
| NSTOR | -2.147 (0.326) | -1.216 (0.585) |
| GAR | 0.867 (0.541) | 0.195 (0.894) |
| AGE | -0.016 (0.713) | -0.035 (0.430) |
| LnLOTSZ | 6.461 ($6.6\ 10^{-7}$) | 5.815 ($8.7\ 10^{-6}$) |
| LnSQFT | 9.881 (0.004) | 8.724 (0.010) |
| PATIO1 | 3.652 (0.093) | 2.091 (0.405) |
| FIREPL1 | 5.637 (0.002) | 4.526 (0.022) |
| AC1 | 6.233 (0.001) | 4.956 (0.026) |
| $\lambda$ | 0.934 (0.001) | 0.930 ($2.210^{-4}$) |
| AIC | 1469.2 | 1466.6 |

Table: SEM Regression coefficients estimates, p-value significance in parenthesis

# Application 3: Non-Spatial Data (Bank Salaries)

**Dataset Description:**

- 474 employees hired by a Midwestern bank.
- Dependent variable $Y$: `log(salary)`.
- Covariates $X$: `log(salbegin)`, `jobtime`, `prevexp`, `educ`, `sex`, `minority`.
- **Objective:** Apply spatial econometrics to non-geographically referenced datasets!

## Synthetic Graph Construction

1. Clustering on quantitative covariates (Kohonen $4 \times 3$, GNG 9 clusters).
2. Distance matrix between prototypes.
3. Construction of an artificial spatial weight matrix $W$.

# Results: OLS vs Spatial SAR Comparison

| **Variable** | **OLS Model** | **SAR $W_{1/d_K}$ (Kohonen)** | **SAR $W_{1/d_{GNG}}$ (GNG)** |
|---|---|---|---|
| (Intercept) | 1.9788 | 0.3204 (0.0001) | 0.4092 ($1.3 \cdot 10^{-5}$) |
| logsalbegin | 0.8996 ($< 2 \cdot 10^{-16}$) | 0.8354 ($< 2 \cdot 10^{-16}$) | 0.8433 ($< 2 \cdot 10^{-16}$) |
| jobtime | 0.0051 ($3.9 \cdot 10^{-11}$) | 0.0045 ($3.9 \cdot 10^{-9}$) | 0.0049 ($8.8 \cdot 10^{-11}$) |
| prevexp | -0.0005 ($1.5 \cdot 10^{-12}$) | -0.0007 ($5.0 \cdot 10^{-11}$) | -0.0007 ($6.5 \cdot 10^{-13}$) |
| educ | 0.0098 (0.0112) | 0.0109 (0.0055) | 0.0113 (0.0034) |
| minority Yes | -0.0267 (0.1537) | -0.0288 (0.1303) | -0.0321 (0.0867) |
| sex Fem | -0.0601 (0.0012) | -0.0551 (0.0041) | -0.0656 (0.0005) |
| $\rho$ **(Autocorrelation)** | — | **-0.0219** (0.0041) | **-0.0257** (0.0045) |
| **AIC** | **-361.84** | **-368.09** | **-367.91** |

Table: Model comparison for bank salaries. Neural spatial modeling significantly improves AIC compared to OLS.

# Application 4: Covid-19 in Habana (dataset 2022)

## Epidemiological & Health

- **Cases_Pop**: Number of cases / population
- **Prop_Hip**: Proportion with hypertension
- **Prop_Diab**: Proportion with diabetes
- **Prop_Card**: Proportion with ischemic heart disease
- **Prop_Epoc**: Proportion with COPD
- **Prop_Asthma**: Proportion with asthma
- **Vaccination2**: Vaccination period in health area

## Demographics & Housing

- **Prop_Under14**: Proportion under 14 years old
- **Prop_Over60**: Proportion over 60 years old
- **Ageing_index**: Ageing index
- **Overcrowding**: Overcrowding (persons / room)
- **Avg_PcVivT**: % modern private housing (types I & II)
- **Avg_GradxE**: Average education level

## Healthcare Access & Services

- **Dist_to_HCS**: Distance to primary health care centers
- **FDO_Density**: Family physician offices density
- **Pop_at_FDO**: Avg. population per family physician office
- **Pop60oHper**: Pop. over 60 / health personnel count

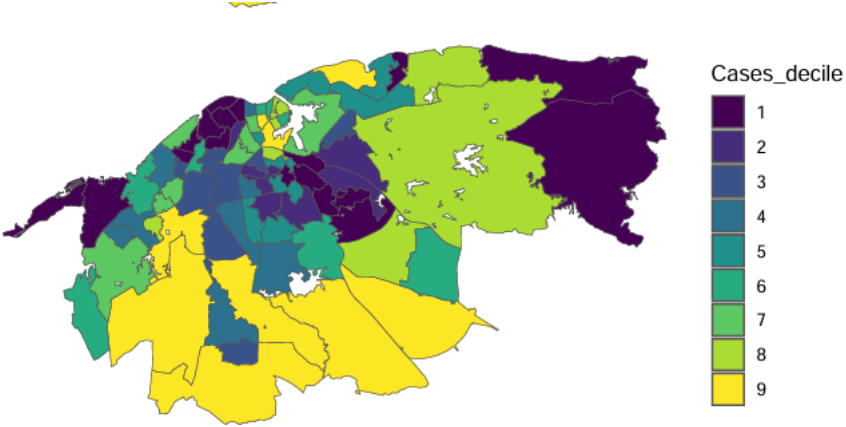
## Geography & Infrastructure

- **QtyNBprec**: Number of precarious neighborhoods
- **Agglomeration**: Number of crowd-attracting places
- **Pop_Dens**: Population density
- **Health_Area**: Health area
- **Municipality**: Municipality

# Application 4: Covid-19 in Habana (dataset 2022)

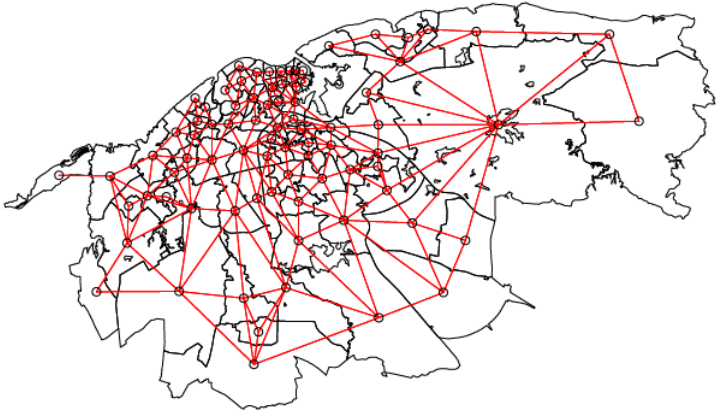
1. Many issues in the dataset:
   - We noticed that the maximum value of the Prop Under14 variable is 1.3, though it represents a proportion of the population.
   - There are many similar inconsistencies in the data.
   - To date, we still do not have a reliable data table.
2. A huge number of covariates, high multicolinearity

# Application 4: Covid-19 in Habana



# Application 4: Covid-19 in Habana

It makes no sense to use geographical distances here.



# Moran's index

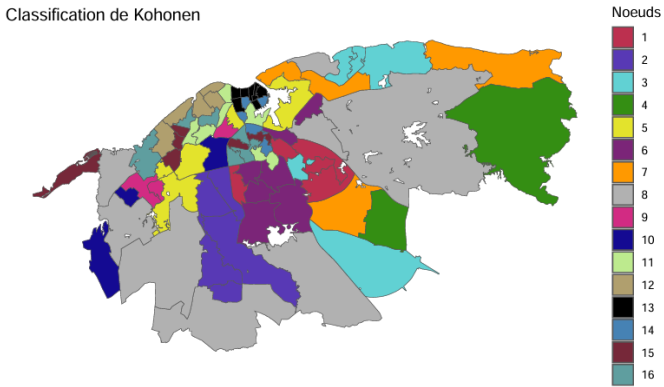
**Queen neighborhood including distances:** this configuration is a combination of the first two, calculated by taking the element-wise product of the distance and Queen neighborhood matrices.

Table: Calculation of Moran's I

| | Moran's I | Expected | Variance | Statistic | p-value |
|---|---|---|---|---|---|
| **Complete graph** | | | | | |
| listw_D | 0.025 | -0.012 | 0.000 | 2.695 | 0.007 |
| listw_racD | 0.005 | -0.012 | 0.000 | 3.354 | 0.001 |
| **Queen neighborhood** | | | | | |
| listw_Q | 0.249 | -0.013 | 0.003 | 4.492 | 0.000 |
| **Queen neighborhood including distances** | | | | | |
| listw_QD | 0.138 | -0.013 | 0.004 | 2.280 | 0.023 |
| listw_racQD | 0.183 | -0.013 | 0.003 | 3.375 | 0.001 |

# Kohonen map

Hexagonal grid $4 \times 4$




- 13: Pop. density, overcrowding (persons/room), Avg. pop per physician, % modern private housing
- 8: Distance to primary health care centers

# Moran's index

| | Moran's I | Expected | Variance | Statistic | p-value |
|---|---|---|---|---|---|
| **Kohonen and Queen neighborhood** | | | | | |
| listw_QK | 0.411 | -0.013 | 0.005 | 5.923 | 0.000 |
| listw_racQK | 0.325 | -0.013 | 0.004 | 5.477 | 0.000 |
| **Kohonen and Queen neighborhood including distances** | | | | | |
| listw_QDK | 0.180 | -0.013 | 0.008 | 2.197 | 0.028 |
| listw_racQDK | 0.225 | -0.013 | 0.004 | 3.820 | 0.000 |

# Spatial Regression Results: SEM vs. SAR Models

- **Spatial Error Model (SEM QK):**
  - Spatial Parameter ($\lambda$) positive and significant.
  - **Positive & Significant:** Distance to health services.
  - **Negative & Significant:** Population over 60 relative to healthcare personnel and percentage of modern private housing.
- **Spatial Autoregressive Model (SAR QK):**
  - Spatial Parameter $\rho$ positive and significant.
  - **Positive & Significant:** Distance to health services and population density.
  - **Negative & Significant:** Ratio `Pop60Hper` and modern housing.
  - Needs impacts study

# Conclusion & Key Contributions

## Summary of Contributions

1. **Novel Neighbourhood Metric:** Incorporating neural clustering distances ($W_K$, $W_{GNG}$) outperforms traditional inverse geographical distance.
2. **Solving the $W_2$ Choice:** Provides a natural, theoretically sound second weight matrix $W_2$ for general spatial models (SAC/SARAR).
3. **Generalization to Non-Spatial Data:** Enables spatial econometric methods on any dataset without geographic coordinates via synthetic graphs.

## Perspectives for Future Work

- Work on a reliable dataset !
- Investigation of other clustering techniques
- Multicollinearity

# Thank you for your attention!

Questions & Discussion