Local summary characteristics for marked spatial point processes with composition-valued marks

Master's Thesis Project

Clemens Baldzuhn

Supervisor: Dr. Matthias Eckardt

June 17, 2025

Table of contents

Introduction

- Georeferenced data becomes rapidly more available which brings with it the need to accurately model it statistically.
- **Spatial point processes** can model point data without spatially discretizing it.
- **Marked spatial point processes** can describe point patterns with additional information attached to the points.
  - e.g. Players positions on the field and distance-covered at a specific time point of a soccer match.
- Marks can be of many types, whereas this thesis focuses on **composition-valued** marks.
  - Compositional data is representable as pie chart. E.g. proportion of time spent walking vs. running vs. sprinting in the given example.
- Marked spatial point processes are analysed using **mark summary characteristics**, which Eckardt et al. (2025) recently proposed for composition-valued marks.

Introduction

Aim of the thesis

- Mark summary characteristics are typically computed globally by averaging over all points, thereby overshadowing spatial heterogeneity in mark distributions (Eckardt & Moradi, 2024).
- To overcome this, we develop **local mark summary characteristics for composition-valued marks** that can be from the perspective of individual points.
- The methods methods are tested in several simulation scenarios and applied to real-world data on economic sector compositions in Castile-La Mancha.

Marked spatial point process

We consider a realization from a marked spatial point process on $\mathbb{R}^2 \times \mathbb{S}^D$ (point in the plane, marks in simplex) denoted by

$$X = \{(x_i, m_i)\}_{i=1}^{n}.$$

The process is assumed to be stationary such that

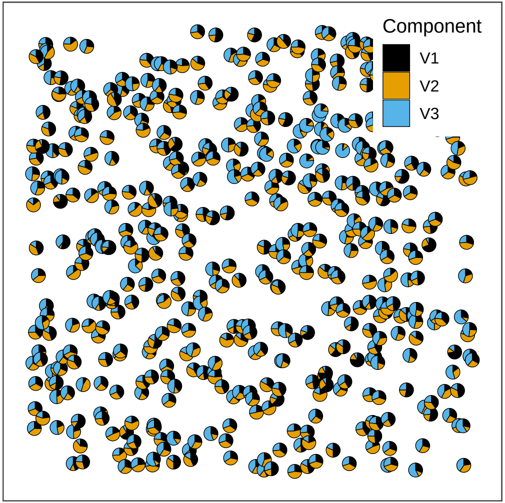
$$\{(x, m)\} \overset{d}{=} \{(x + s, m)\}.$$

The expected number of points with locations in B and marks in L

$$\begin{aligned} \Lambda(B \times L) &= \mathbb{E}[N(B \times L)] \\ &= \mathbb{E}\left[\sum_{(x,m)\in X} \mathbb{1}_{B\times L}(x, m)\right], \end{aligned}$$

is also a measure since it is non-negative and countably additive over disjoint measurable subsets from X.

Marked spatial point process



Composition-valued marks

Marks are D-part compositions, which are essentially real-valued vectors

$$m_i = (m_i^1, m_i^2, \ldots, m_i^D) \in \mathbb{S}^D.$$

The non-standard geometry of the Simplex however forbids applying usual statistics.

Simple transformations ψ such as the **centered-logratio transformation** (clr) exist to transform marks into standard Euclidean space.

$$\psi_j(m_i) = \log(m_i^j / g(m)), \quad j = 1, \ldots, D, \quad g(m) := \left(\prod_{j=1}^{D} m^j \right)^{1/D},$$

where each composition-valued mark component is relative to the geometric mean.

The clr-transformation is isometric but not injective and hence not isomorph since $\sum_{j=1}^{D} \mathrm{clr}(m_i^j) = 0$.

Composition-valued marks

Applying the transformations to the marks:

$$\psi(m_i) = (m_i^{\psi,1}, m_i^{\psi,2}, m_i^{\psi,D})$$

Mark summary characteristics are formulated on the basis of **test functions of marks** to describe numerical behaviours of marks. We will express them as

$$\mathbf{t}_f^{\psi,jl}(m_\circ^{\psi,j}, m_k^{\psi,l})$$

Summary characteristics for composition-valued marks

Global case: Mark summary characteristics describe some expected numerical relationship between marks of points within a fixed inter point distance.

$$\mathbb{E}_\circ \left[\mathbf{t}_f^{\psi,jl}(m_\circ^{\psi,j}, m_k^{\psi,l}) \mid ||x_\circ - x_k|| = r \right] \tag{1}$$

The expectation is conditional on

1. $x_\circ$ and x_k indeed being present in X, and
2. $x_\circ$ and x_k having an interpoint distance r.

The subscript $\circ$ indicates that $x_\circ$ is in a *typical* location, such that (1) represents an *average over marks of all points* with specific interpoint distance.

Summary characteristics for composition-valued marks

Local case: Constructed by fixing a specific point x_i and some point x_k.

$$\mathbb{E}_i \left[\mathbf{t}_{f,i}^{\psi,jl}(m_i^{\psi,j}, m_k^{\psi,l}) \mid ||x_i - x_k|| = r \right]$$

The condition again ensures that points x_i and x_k are indeed points in X and within distance r.

The additional indices emphasize taking the *perspective of a specific point* x_i.

This gives the expected numerical mark behaviour with points in x_i's r-distant vicinity.

Local mark summary characteristics

$\mathbf{t}_{1,i}^{\psi,jl} = m_i^{\psi,j} m_k^{\psi,l}$ corresponds to a local **unnormalized mark correlation function** ($\tau_i^{\psi,jl}$) describing the association of marks. (Stoyan, 1984).

$\mathbf{t}_{2,i}^{\psi,jl} = 0.5(m_i^{\psi,j} - m_k^{\psi,l})^2$ corresponds to a local **mark variogram** ($\gamma_i^{\psi,jl}$) describing the similarity/variability of marks. (Cressie, 1993).

Normalization factors for both can be obtained by computing the expected test function under mark independence, i.e. for $r \to \infty$.

Interpretation for normalized versions: Values larger than one indicate that marks within a given distance are

1. positively associated with each other more than would be expected under random labelling/mark independence (τ_i)
2. less similar than expected (γ_i)

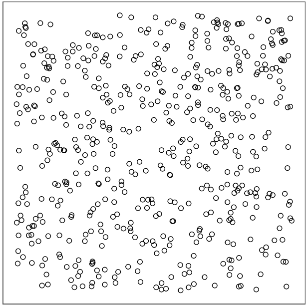
Simulation study

The local summary characteristics for composition-valued marks are tested against their global analogues on

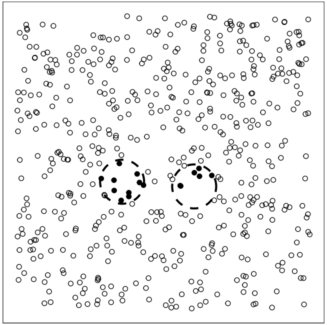
- the frequency with which both erroneously detect mark dependence (type 1 errors)
- the extent to which they are able to capture actually present mark dependence (statistical power)

Result based on 500 patterns in 4 simulations scenarios. Significance estimates obtained using global envelope tests (Myllymäki et al., 2024).

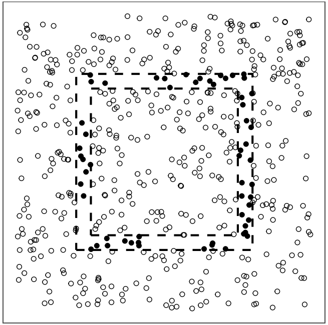
Simulation Scenarios



(a) Scenario I: No dependence

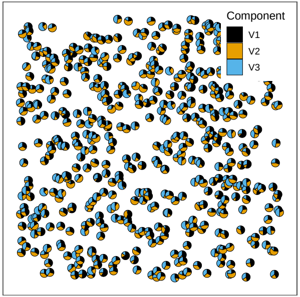


(b) Scenario II: Clustered dependence

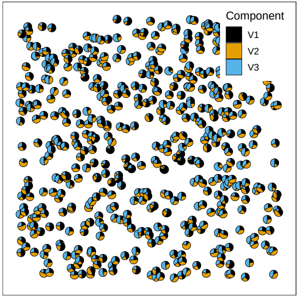


(c) Scenario IV: Regular dependence

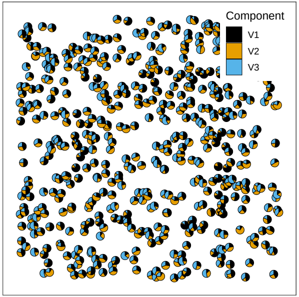
Simulation scenarios



(a) Scenario I

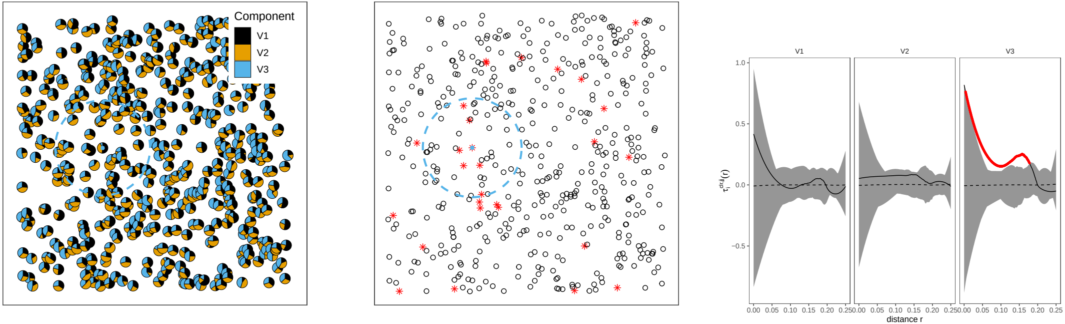


(b) Scenario II



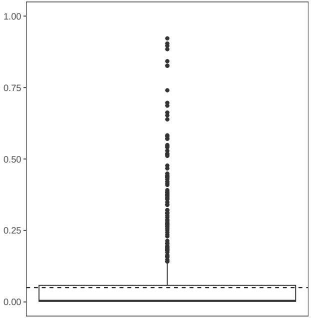
(c) Scenario IV

Scenario 1

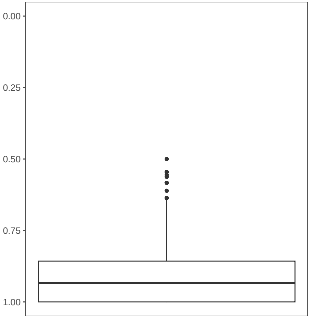


(a) (b) (c)

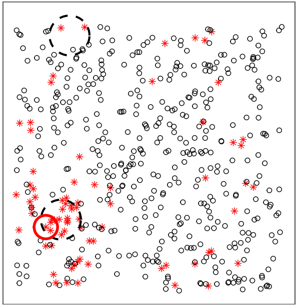
Scenario 2



(a) p-values obtained from GET tests on 500 simulated patterns in global analysis



(b) Fraction of correctly detected points in local analysis



(c)

Results

Type I Errors: Both global and local methods showed comparable, expected Type I error rates (around 5%) when no mark dependence was present.

Clustered Dependence Power: Local statistics significantly outperformed global ones in detecting clustered mark dependence, with 90-91% correct detection versus 70-74% for global methods.

Regular Dependence Power: Local statistics performed less effectively (46.7%) than global ones (69.2%) for regular mark dependence patterns, highlighting a limitation in this scenario.

Application

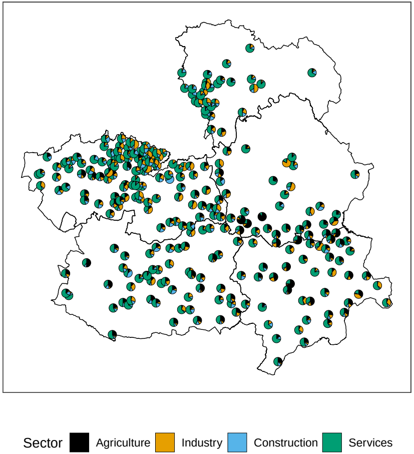
Employing the introduced statistics, we analyze economic sector compositions for 278 municipalities in Castile-La Mancha.

The locations of municipalities in Castile-La Mancha can be understood as a realization from a Poisson Point process.

Marks are obtained from the number of workers per economic sector (Gobierno Regional de Castilla-La Mancha, 2025a).

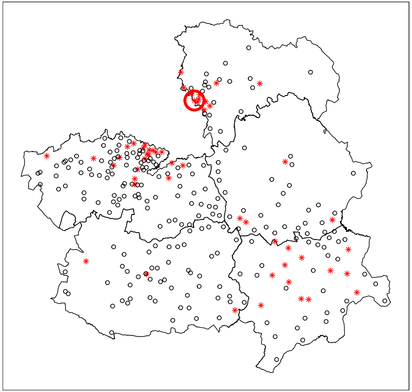
Data

(a)

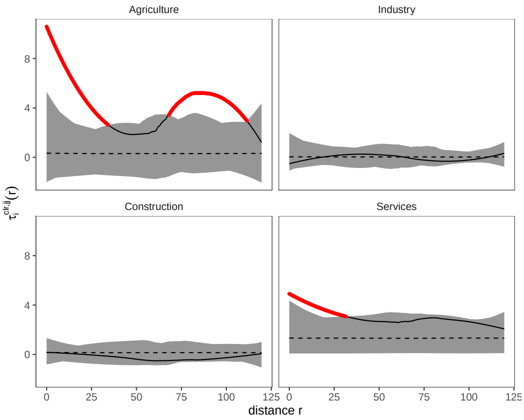


(b)

Mark associations I

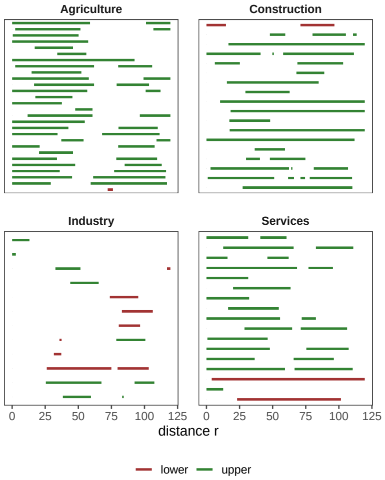


(a) Red stars: Signficant associations with surrounding marks (w.r.t. all economic sectors)



(b)

Mark associations II



(a)

(b)

Discussion

- Results emphasize analytical depth obtained from employing local summary statistics
- Future applications should extend to local cross-relation mark summary characteristics to investigate synergies/interferences between economic sectors (Ahlfeldt & Pietrostefani, 2019).
- Local summary characteristics should be generalized to non-stationary ground processes.

References I

Ahlfeldt, G. M., & Pietrostefani, E. (2019).The Economic Effects of Density: A Synthesis. *Journal of Urban Economics*, *111*, 93–107. Retrieved June 11, 2025, from https://www.sciencedirect.com/science/article/pii/S0094119019300282

Cressie, N. (1993). *Statistics for Spatial Data*. Wiley. https://doi.org/10.1002/9781119115151

Eckardt, M., & Moradi, M. (2024).Local Indicators of Mark Association for Spatial Marked Point Processes. https://doi.org/10.48550/arXiv.2411.01065

Eckardt, M., Myllymäki, M., & Greven, S. (2025).On Spatial Point Processes with Composition-Valued Marks. https://doi.org/10.48550/arXiv.2501.16049

Gobierno Regional de Castilla-La Mancha. (2025a). *Portal Estadistica de Castilla-La Mancha*. Afiliacion a uttimo dia de mes por sectores económicos y municipios. https://difusion.jccm.es/wds/TableViewer/tableView.aspx?ReportId=49493

References II

Gobierno Regional de Castilla-La Mancha. (2025b). *Portal Estadistica de Castilla-La Mancha*. Descargas de Datos Geográficos Abiertos. https://datos-abiertos-castillalamancha.opendata.arcgis.com/search?tags=OrdenacionTerritorial

Myllymäki, M., Mrkvicka, T., Kuronen, M., Dvorak, J., Grabarnik, P., Hahn, U., Rost, M., & Seijo, H. (2024). *GET: Global Envelopes* (Version 1.0-4). https://cran.r-project.org/web/packages/GET/index.html

Stoyan, D. (1984).On Correlations of Marked Point Processes. *Mathematische Nachrichten*, *116*(1), 197–207. https://doi.org/10.1002/mana.19841160115