The Annals of Regional Science manuscript No.
(will be inserted by the editor)

Measuring spatial dispersion: exact results on the variance of random spatial distributions

Pablo Jensen · **Julien Michel**

Received: date / Accepted: date

Abstract Measuring the spatial distribution of locations of many entities (trees, atoms, economic activities, etc.), and, more precisely, the deviations from purely random configurations, is a powerful method to unravel their underlying interactions. Several coefficients have been developed in the past to quantify the possible deviations. It is important to quantify the variances of the coefficients for random distributions, to ascertain the statistical significance of an empirical deviation. By lack of a proper analytical expression, the significance is usually obtained by simulating many random configurations by Monte Carlo simulations. In the present paper, we present an *exact* analytical expression for the variance of several spatial coefficients for random distributions, and we *rigorously* show that these distributions asymptotically follow a Normal law. These two results eliminate the need for cumbersome Monte Carlo simulations. They also allow to understand qualitatively the main factors that may change the variance: number of sites, spatial inhomogeneity, etc.

Keywords Spatial distribution · Point processes · Statistical significativity · Localisation

JEL C19 · R12 · L81

1 Introduction

Measuring the spatial distribution of industries [Hoo37], atoms [EB03], trees [WPF96] or retail stores [HG84, Col06, Jen06] is a powerful method to understand the underlying mechanisms of their interactions. Several methods have been developed in the past to quantify

Pablo Jensen
Université de Lyon
(a) LET-CNRS, Université Lyon-2, 69007 Lyon, FRANCE
(b) Laboratoire de Physique, Ecole Normale Supérieure de Lyon et CNRS, 69007 Lyon, FRANCE
(c) Institut des Systèmes Complexes Rhône-Alpes, IXXI-CNRS, 69007 Lyon, FRANCE
E-mail: Pablo.Jensen@ens-lyon.fr

Julien Michel
Université de Lyon
(d) Unité de Mathématiques Pures et Appliquées-UMR 5669, ENS Lyon, 46 allée d'Italie, F-69364 Lyon Cedex 07, FRANCE
E-mail: Julien.Michel@umpa.ens-lyon.fr

the deviations of the empirical distributions from *purely random distributions*, supposed to correspond to the non-interacting case [Rip76, Bes77, EG97, MS99]. Recently, a method originally developed by G. Duranton and H. Overman [DO05], later modified by Marcon and Puech [MP07] has been proposed. Its main interest is that it takes as reference for the underlying space not a homogeneous one as for the former methods [Rip76, Bes77, EG97, MS99], but the overall spatial distribution of sites, thus automatically taking into account the many inhomogeneities of the actual geographical space. For instance, retail stores are inhomogeneously distributed because of rivers, mountains or specific town regulations (parks, pure residential zones, etc.). Therefore, it is interesting to take this inhomogeneous distribution as the reference when testing the random distribution of, for instance, bakeries, in town. Furthermore, by using precise location data (x and y coordinates), this method avoids all the well-known contiguity problems, summarized in the 'modifiable areal unit problem' [YK50, Unw96, Ope84, BCL07]. However, the method has two main drawbacks:

1. the need of precise location data (*i.e.* x, y coordinates, and not only knowing that a site belongs to a given geographical area),
2. the need for Monte Carlo simulations in order to compute the statistical significance of the deviations from a random distribution.

Point (1) is probably going to be less crucial as precisely spatialized data becomes more common. Moreover, it can be argued that, when only region-type data exists, it can be more convenient to locate all the sites at the region centroid and then apply the 'continuous' method, thus avoiding contiguity problems.

This paper solves point (2), and, more generally, gives analytic formulas to compute the variance of some characteristics of purely random distributions of points (meaning non interacting distributions). This is all the more important since the variance is needed to ascertain the statistical significance of a deviation from the reference value of the spatial distribution. Lacking an analytical expression for the variance, the usual method to compute the statistical significance of a deviation consists in generating random distributions by Monte Carlo simulations, counting the proportion of distributions that deviate more than the measured distribution. However, Monte Carlo simulations can be cumbersome to implement and furthermore this method can be prohibitively time consuming for large samples, since typically many thousand runs are necessary to compute deviations precisely.

The new indices defined in this article present the great advantage, compared to the classical indices, of having a reference value constantly equal to 1 (no additional computation by Monte Carlo simulation of this reference is thus needed). The deviation from this reference of purely random configurations is clearly observed either for clustered or self-excluding configurations, and those behaviors are readily obtained by comparing the indices with 1: significantly greater than the reference 1 means clustering, whereas lesser than the reference means exclusion. The significancy levels for this discrimination are determined thanks to the knowledge of the variances of the indices. The indices introduced cover both the case of discrete locations, and of continuous locations with respect to an *a priori* density measure. Our main results obtained are theoretical values for the variance of all the indices introduced, giving an easy access to their actual numerical computation. This numerical computation is much faster than the classical Monte Carlo simulations that were used up to now in the litterature. Thus, testing deviation from randomness for a spatial configuration of millions of stores becomes practically feasible. Furthermore, the asymptotic normality proved in the appendix ensures that the variance suffices to calculate the confidence intervals.

The paper is organized as follows: in section 2 we define the cumulative coefficient [MP07] measuring the spatial dispersion or aggregation of locations (stores, etc.) in a discrete setting. In section 3 we derive explicit values for the variance of the discrete cumulative coefficients and give an example on the actual locations of bakeries in the city of Lyon. Section 4 is devoted to a more theoretical derivation of the variance in the continuous setting. Finally, in section 5, we extend our calculations to other spatial coefficients, such as a differential index and the widely used Duranton-Overman index [DO05].
The exact value of the variance for the discrete case is given in proposition 1 for the *inter*-coefficient measuring the dependency of the locations of a certain type of stores with respect to another type of fixed stores. Proposition 2 gives the value for the variance of the *intra*-coefficient measuring the dependency between the same type of stores. The testing procedure is detailed in proposition 3. The continuous version is treated in section 4, and the computation of the variances in this case is performed in propositions 4 and 5. Theorems 1 and 2 state central limit theorems justifying the normal approximation for large numbers of stores. Propositions 6 and 7 give the value of the variance of respectively the Differential coefficients and of the Duranton-Overman indicator.

List of Figures

List of Tables

2 Discrete setting for the characterisation of spatial dispersion and aggregation

The indicators that are studied here deal with the problem of quantifying deviations of empirical distribution of points from purely random and non-interacting distributions. One can be interested in the interaction of a set of points between themselves, or with some other set of points. From now on we shall work with two different types of points: A and B. We define two indicators, refered to as respectively the *inter* and *intra*-coefficients [MP07], to characterize the (cumulative) spatial interaction between sites closer than a distance r, the *inter*-coefficient describes the type of interactions of fixed A points with random B points, whereas *intra*-coefficient is intended to measure the independence between points of type A. One can also work with indicators characterizing the (differential) spatial distributions between distances r and $r+\delta r$ (with $\delta r \ll r$) [DO05]. Those differential coefficients are potentially more sensitive to spatial variations of the distributions because they do not integrate features from 0 to r. We shall start by calculating the variance of the cumulative coefficient and then extend our results to other quantifiers of spatial distributions.

In a discrete setting, stores (points) are located randomly on a discrete (finite) set T of fixed points. We shall use the following definitions and notations:

- there are N_t sites, of which N_A sites are of type A, and N_B sites are of type B,
- for any site S, we define:

$N_t(S,r)$	the total number of sites distinct from S that are at a distance lesser than r of site S,
$N_A(S,r)$	the number of A sites in this same region
$N_B(S,r)$	the number of B sites in this same region.

Note that site S is *never* counted in those quantities, whatever its state.

Remark 1 The notation $N_t(D)$ (resp. $N_A(D)$ and $N_B(D)$) will denote the total (resp. A and B) number of sites in a subset D of T. Thus for instance $N_A(S,r)$ stands for $N_A(B(S,r)\setminus\{S\})$, where $B(S,r)$ denotes the ball centered at S with radius r.

In this discrete model, the locations of stores A and B are distributed over the total number of possible sites, with mutual exclusion at a same site. Therefore, the geographical characteristics of the studied area are carried by the actual locations of those possible N_t sites.

Remark 2 The coefficients that we introduce depend on the reference distance r, however we shall drop this dependency in the notations, unless when strictly necessary.

In the following two subsections we define the *inter* and *intra*-coefficients, with the following goal: both coefficients must be easy to compute on actual data, and they should be easily compared to a reference value associated to a random distribution of points. The computation of their variance shall then give an easy to implement *testing hypothesis* for this randomness by a standard argument of confidence interval based on Chebyshev's inequality [YK50]. This testing procedure will be detailed in subsection 3.4.

2.1 *inter*-coefficient

In order to quantify the dependency between two different types of points, we set the following context: the set T has a fixed subset of N_A stores of type A, and the distribution of the subset $\{B_i, i=1...N_B\}$ of type B stores is assumed to be uniform on the set of subsets of cardinal N_B of $T\setminus\{A_1,\ldots,A_{N_A}\}$: this is equivalent to an urn model with N_B draws with no replacement in an urn of cardinal N_t-N_A. The presence of a point of type A at those locations, under this reference random hypothesis, should not modify (in average) the density of type B stores: the local B spatial concentration $(N_B(A_i,r))\,/\,(N_t(A_i,r)-N_A(A_i,r))$ should be close (in average) to the concentration over the whole town, $(N_B)\,/\,(N_t - N_A)$. We define the *inter*-coefficient as

$$a_{AB} = \frac{N_t - N_A}{N_A N_B} \sum_{i=1}^{N_A} \frac{N_B(A_i, r)}{N_t(A_i, r) - N_A(A_i, r)} \tag{1}$$

where $N_A(A_i, r)$, $N_B(A_i, r)$ and $N_t(A_i, r)$ are respectively the A, B and total number of points in the r-neighborhood of point A_i (not counting A_i), *i.e.* points at a distance smaller than r. In this definition, the right hand side may contain fractions with zero at the numerator and denominator: those fractions $0/0$ are taken as equal to 1.
It is straightforward to check that

Lemma 1 *For all $r > 0$, we have $E[a_{AB}] = 1$.*

We can deduce a qualitative behaviour in the following sense: if the observed value of the *inter*-coefficient is greater than 1, we may deduce that A stores have a tendency to attract B stores, whereas lower values mean a rejection tendency.

2.2 *intra*-coefficient

Let us assume that we are interested in the distribution of N_A points in the set T, represented by the subset $\{A_i, i = 1...N_A\} \subset T$. The reference law for this set, called *pure random distribution*, is that this subset is uniformely chosen at random from the set of all subsets of cardinal N_A of T: this is equivalent to an urn model with N_A draws with no replacement in an urn of cardinal N_t.

Intuitively, under this (random) reference law, the local concentration represented by the ratio $N_A(A_i, r)/N_t(A_i, r)$ of stores of type A around a given store of type A should, in average, not depend on the presence of this last store, and should thus be (almost) equal to the global concentration N_A/N_t, this leads us to introduce the following *intra*-coefficient:

$$a_{AA} = \frac{N_t - 1}{N_A(N_A - 1)} \sum_{i=1}^{N_A} \frac{N_A(A_i, r)}{N_t(A_i, r)}. \tag{2}$$

In this definition, the fraction $0/0$ is still taken as equal to 1 in the right hand term.
Under the *pure randomness hypothesis*, it is straightforward to check that the average of this coefficient is equal to 1:

Lemma 2 *For all $r > 0$, we have $E[a_{AA}] = 1$.*

We also deduce a qualitative behaviour in the following sense: if the observed value of the *intra*-coefficient is greater than 1, we may deduce that A stores tend to aggregate, whereas lower values indicate a dispersion tendency.

3 Computation of the variance in the discrete setting, inference for the stores data of the city of Lyon

The computation of the variances of the *inter* and *intra*-coefficients does not contain many mathematical difficulties, the most important feature is the fact that in the computation of the second moment of this coefficient, the possible overlaps of neighborhoods yields a loss of independence.

3.1 *Inter* covariance

Let us recall that there are N_A particular (fixed) sites of type A, and that sites B are taken randomly. The r-neighborhoods of the fixed positions $A_1, \dots, A_{N_A}$ may intersect. We shall denote by $C^r_{i,j} = B(A_i, r) \cap B(A_j, r)$ the intersection of the two balls of centers A_i and A_j and radius r. In the computation of the second order moment of a_{AB} we shall use the simplified notations:

- $N^i_B = N_B(A_i, r)$,
- N^{ij}_B the number of B stores in $C^r_{i,j}$,
- $N^{i\setminus j}_B$ the number of B stores in $B(A_i, r) \setminus C^r_{i,j}$.

Using those notations, we may write:

$$\begin{aligned}
a^2_{AB} &= \left(\frac{N_t - N_A}{N_A N_B} \sum_{i=1}^{N_A} \frac{N^i_B}{N_t(A_i, r) - N_A(A_i, r)} \right)^2, && (3)\\
&= \left(\frac{N_t - N_A}{N_A N_B} \right)^2 \left\{ \sum_{i=1}^{N_A} \left(\frac{N^i_B}{N_t(A_i, r) - N_A(A_i, r)} \right)^2 + \right.\\
&\left. \sum_{i \neq j} \frac{N^i_B}{N_t(A_i, r) - N_A(A_i, r)} \frac{N^j_B}{N_t(A_j, r) - N_A(A_j, r)} \right\}, && (4)\\
&= \left(\frac{N_t - N_A}{N_A N_B} \right)^2 \left\{ \sum_{i=1}^{N_A} \left(\frac{N^i_B}{N_t(A_i, r) - N_A(A_i, r)} \right)^2 + \right.\\
&\left. \sum_{i \neq j} \frac{N^{i\setminus j}_B N^{j\setminus i}_B + N^{i\setminus j}_B N^{ij}_B + N^{ij}_B N^{j\setminus i}_B + \left(N^{ij}_B\right)^2}{(N_t(A_i, r) - N_A(A_i, r))(N_t(A_j, r) - N_A(A_j, r))} \right\}. && (5)
\end{aligned}$$

All the terms in the right hand side are either squares of numbers of B stores in some region, or products of two such terms in *disjoint* regions: the computation of their expectation is quite easy thanks to the following elementary lemma:

Lemma 3 *Let D be a subset of cardinal d of $T^{\setminus A} = T \setminus \{A_1, \dots, A_{N_A}\}$, and let D' be a disjoint subset of cardinal d' of $T^{\setminus A}$. Under the* pure random hypothesis*, the numbers $N_B(D)$ and $N_B(D')$ of B stores in D and D' have the following properties:*

$$\begin{aligned}
P(N_B(D) = k) &= \binom{d}{k} \binom{N_t - N_A - d}{N_B - k} \Big/ \binom{N_t - N_A}{N_B},\\
E[N_B(D)] &= d \, \frac{N_B}{N_t - N_A},\\
E[N_B(D)^2] &= d(d-1) \, \frac{N_B(N_B - 1)}{(N_t - N_A)(N_t - N_A - 1)} + d \, \frac{N_B}{N_t - N_A},\\
E[N_B(D) N_B(D')] &= dd' \, \frac{N_B(N_B - 1)}{(N_t - N_A)(N_t - N_A - 1)}.
\end{aligned}$$

Remark 3 If we set $p = d/(N_t - N_A)$, the variance $\sigma^2(N_B(D))$ becomes

$$\sigma^2(N_B(D)) = p(1-p) N_B \, \frac{N_t - N_A - N_B}{N_t - N_A - 1}.$$

Using the computations of this lemma, the value of the variance of the *inter*-coefficient follows easily: if we denote by $\langle\cdot\rangle_A$ the average over the A sites,

$$\langle u(A_i)\rangle_A := \frac{1}{N_A}\sum_{i=1}^{N_A} u(A_i), \text{ and } \langle k(A_i,A_j)\rangle_A := \frac{2}{N_A(N_A-1)}\sum_{1\le i<j\le N_A} k(A_i,A_j),$$

we obtain

Proposition 1 *The variance of the* inter-*coefficient is given by*

$$\begin{aligned}\sigma^2(a_{AB}) = &-\frac{N_t-N_A-N_B}{N_B(N_t-N_A-1)}\\ &\left(\frac{N_t-N_A}{N_AN_B}\frac{N_t-N_A-N_B}{N_t-N_A-1}\right)\left\langle\frac{1}{N_t(A_i,r)-N_A(A_i,r)}\right\rangle_A\\ &+\left(\frac{N_t-N_A-N_B}{N_t-N_A-1}\right)\left(\frac{N_t-N_A}{N_B}\right)\left(1-\frac{1}{N_A}\right)\left\langle x_{ij}\right\rangle_A,\end{aligned}$$

where

$$\left\langle x_{ij}\right\rangle_A = \left\langle\frac{N_t(C_{i,j})-N_A(C_{i,j})}{(N_t(A_i,r)-N_A(A_i,r))(N_t(A_j,r)-N_A(A_j,r)}\right\rangle_A.$$

3.2 *Intra* covariance

The computation of the second moment of the *intra*-coefficient is a little more tricky. When computing the second order moment of a_{AA} we are led to deal with square terms,

$$\sum_i\left(\frac{N_A^i}{N_t(A_i)}\right)^2,$$

(with obvious notations), those are clearly treated using the same arguments as for the *inter* case, and cross-products

$$\sum_{i\neq j}\frac{N_A^i}{N_t(A_i)}\frac{N_A^j}{N_t(A_j)}.$$

To compute the average of those terms, it is better to start from the very definition of the average:

$$\left\langle u(A_i,A_j)\right\rangle_t = \frac{1}{\text{total number of configurations}}\sum_{\text{all configurations}} u(A_i,A_j), \tag{6}$$

where the function u depends on both locations A_i and A_j and on the other points of type A:

$$u(A_i,A_j) = \frac{N_A^i}{N_t(A_i)}\frac{N_A^j}{N_t(A_j)}.$$

Now, we can order all the possible A configurations over the N_t sites by grouping those that keep fixed the positions of A_i and A_j. Therefore, eq. (6) becomes:

$$\left\langle u(A_i,A_j)\right\rangle_t = \frac{1}{\text{total number of configurations}}\sum_{s,t\in T}\sum_{\substack{\text{configurations}\\ \text{such that}\\ A_i=s,A_j=t}} u(A_i,A_j). \tag{7}$$

For s and t fixed, the inner sum in eq. (7), once correctly rescaled, can be interpreted as the average of $u(A_i,A_j)$ when N_A-2 sites are randomly chosen out of N_t-2 sites. In other terms, eq. (7) fixes the positions of A_i and A_j and averages over the positions of all the other A's, in order to compute the average by analogy to the *inter* case seen above, with, formally $N_B \equiv N_A-2$. One has only to be cautious to separate the sum in two terms, for which the value of the random variables $N_A^i N_A^j$ product is different: when A_i and A_j are neighbors, one has to add 1 to the random value of both N_A^i and N_A^j:

$$\begin{aligned} N_A^i &= N_A^{i\backslash j} + N_A^{ij} + 1, \\ N_A^j &= N_A^{j\backslash i} + N_A^{ij} + 1, \end{aligned}$$

where $N_A^{i\backslash j}$ denotes the number of A neighbors of site A_i among all points of T that are r-neighbors of A_i and not r-neighbors of A_j. We introduce the following localised average $\langle\cdot\rangle_n$: for any function v defined on couples of points of T, set

$$\left\langle v(T_i,T_j)\right\rangle_n = \frac{2}{N_t(N_t-1)} \sum_{1\le i<j\le N_t} v(T_i,T_j)\mathbf{1}_{d(T_i,T_j)\le r}, \tag{8}$$

where $\mathbf{1}_{d(s,t)\le r}$ is equal to one if and only if s and t are at a distance lesser than r and to 0 otherwise.

We obtain then the following result:

Proposition 2 *The variance of the* intra-*coefficient is the sum of four terms:*

$$\sigma^2(a_{AA}) = \sum_{i=1}^{4} \mathrm{Var}(a_{AA})_i,$$

where

$$\begin{aligned} \mathrm{Var}(a_{AA})_1 &= -\frac{N_t-N_A}{(N_t-2)(N_A-1)}, \\ \mathrm{Var}(a_{AA})_2 &= \frac{(N_t-1)(N_t-N_A)}{N_A(N_A-1)(N_t-2)} \left\langle \frac{1}{N_t^i} \right\rangle_t, \\ \mathrm{Var}(a_{AA})_3 &= \frac{(N_t-1)^2(N_t-N_A)(N_t-N_A-1)}{(N_t-2)(N_t-3)N_A(N_A-1)} \left\langle \frac{1}{N_t^i N_t^j} \right\rangle_n, \\ \mathrm{Var}(a_{AA})_4 &= \frac{(N_t-1)^2(N_t-N_A)(N_A-2)}{(N_t-2)(N_t-3)N_A(N_A-1)} \left\langle x_{ij}\right\rangle_t, \end{aligned}$$

where $\left\langle x_{ij}\right\rangle_t$ is defined analogously to $\left\langle x_{ij}\right\rangle_A$:

$$\left\langle x_{ij}\right\rangle_t = \frac{2}{N_t(N_t-1)} \sum_{1\le i<j\le N_t} \frac{N_t(T_{i,j}^r)}{N_t(T_i,r)N_t(T_j,r)},$$

where $T_{i,j}^r = B(T_i,r)\cap B(T_j,r)\setminus\{T_i,T_j\}$.

3.3 Numerical evidence and comparison with Monte Carlo simulations

We have performed the computation of the theoretical variances for the set of locations T corresponding to the locations of stores in the city of Lyon, and compared the results with approximate values obtained by Monte Carlo simulations. Figures 1 and 2 give the result of this comparison for different categories of Lyon's town retail stores locations.

Figure 1 presents the comparison of the theoretical *inter* variance and of a Monte Carlo simulation in two situations. The red line shows the analytical results for $N_A = 211$ sites, the circles corresponding to the estimates of the variance from simulation results over a million B configurations for different values of N_B ($\langle 1/N_i \rangle_A = 0.116$, $\left\langle x_{ij} \right\rangle_A = 0.00019$). The blue line corresponds to the analytical results for $N_A = 72$ sites, the squares correspond to the estimates from simulation ($\langle 1/N_i \rangle_A = 0.0674$, $\left\langle x_{ij} \right\rangle_A = 0.000081$.). The N_A values ($N_A = 72$ and $N_A = 211$) are somewhat arbitrary, but nevertheless represent typical numbers of stores of a given activity in a town of roughly a million inhabitants.

Fig. 1 Comparison of variances obtained from simulations and proposition 1 for the *inter*-coefficient as a function of N_B (logarithmic scales).

Figure 2 shows the adequation of the analytical *intra* results (circles) and the variance obtained by a Monte Carlo simulation of a million configurations (red line).

Clearly, the computations closely follow the evolutions of the variances as a function of N_A and N_B. For illustration, the average values for Lyon's sites are (for $r = 100m$) : $N_t = 7839$, $\left\langle 1/N_t^i \right\rangle_t = 0.112$, $\left\langle 1/(N_t^i N_t^j) \right\rangle_n = 1.29\ 10^{-5}$ and $\left\langle x_{ij} \right\rangle_t = 1.19\ 10^{-4}$.

The calculation times for the variances of the 53 retail activities (*i.e.* approximately 2000 terms similar to the points shown in figures 1 and 2) are: 21 seconds with the exact formulas, to be compared with 50 minutes for $10,000$ Monte Carlo simulations.

Those simulations still show some deviations larger than 5% from the exact variance values (mean absolute deviation: 2%). Note that these values are given to allow comparison of the two calculation times, but their absolute values are far from optimal. Standard computation tricks such as R-trees [MNPT05] allow a drastic reduction of calculation times for the exact computation of the variances by computing distances between sites only on sites which are close enough. This leads to a calculation time of about 1 second for 7841 sites with the exact formulas, and, more importantly, a less steep increase with the number of sites. Therefore, the variances with millions of sites can be calculated in a few hours.

Fig. 2 Comparison of variances obtained from simulations and proposition 2 as a function of N_A (linear scales).

3.4 Testing the pure randomness hypothesis

The analytical expressions for the variance of the *inter* and *intra*-coefficients is the major tool in defining easily a test for the randomness hypothesis. Indeed if we define H_0^B to be the following hypothesis: *the locations of B sites are purely random*, and H_0^A the hypothesis *the locations of A sites are purely random*, then we have by Chebyshev's inequality the following results:

Proposition 3 *Let α be a positive (small) number in $(0,1)$. Under hypothesis H_0^B, for any configuration of sites A, we have:*

$$P\left(|a_{AB}-1| \leq q_\alpha^{AB}\right) \geq 1-\alpha, \tag{9}$$

where $q_\alpha^{AB} = \sqrt{\sigma^2(a_{AB})/\alpha}$.
Under hypothesis H_0^A we have

$$P\left(|a_{AA}-1| \leq q_\alpha^{AA}\right) \geq 1-\alpha, \tag{10}$$

where $q_\alpha^{AA} = \sqrt{\sigma^2(a_{AA})/\alpha}$.
This implies that $[1-q_\alpha^{AB}, 1+q_\alpha^{AB}]$ *is a confidence interval with level (at least)* $1-\alpha$ *for* a_{AB} *under hypothesis* H_0^B *(id. for* a_{AA}*).*

Under the alternative hypotheses in both cases it can be shown that the *inter* and *intra*-coefficients have a distinct behavior (this will be highlighted in the next subsection): the expected value of those coefficients is no longer equal to 1 (higher values correspond to clustering whereas lower values correspond to exclusion).
Thus we may formulate the following testing procedure:

– If $|a_{AB}-1| > q_\alpha^{AB}$ reject hypothesis H_0^B.
– If $|a_{AA}-1| > q_\alpha^{AA}$ reject hypothesis H_0^A.

Otherwise accept the hypothesis.

This testing procedure is illustrated in figure 3: the *intra*-coefficient is computed for different values of r (circles), and the vertical bars correspond to the (lower) half confidence interval centered at 1 with level at least $1-\alpha = 0.95$.

Fig. 3 Plot of the *intra*-coefficient for bakeries in the city of Lyon with respect to r.

The figure 3 shows the practical importance of variance calculations for economic interpretations of the data. Although a_{AA} remains well below the reference value (*i.e.* 1), bakeries are significantly dispersed only until $150m$ (the hypothesis H_0^A is rejected). For longer distances, their spatial locations approach a random pattern: the hypothesis H_0^A is not rejected for large values of r.

3.5 Clustering and exclusion: examples and numerical evidence of the deviation from the pure random case

The theoretical computations in the case of the pure random hypothesis gives the reference model, however we need to show that different situations yield (radically) different values for the *inter* and *intra*-coefficients. This is also illustrated in [MP07], p. 15–17.

For the *inter* case it is straightforward to generate random configurations showing clustering or exclusion around sites of type A: let us indeed consider for the B sites for instance a pure random configuration outside the r_0-neighborhoods of the A points (if the total number of sites satisfying this condition is larger than N_B), then it is clear that by construction $a_{AB} = 0$ for all $r \leq r_0$. On the contrary if we concentrate (randomly) the B points around the A points, we may obtain a coefficient (largely) greater than 1.

We shall detail the generation of such random configurations for the *intra*-coefficient, showing clustering or exclusion properties. Let us consider the square lattice $[-N,N]^2$, and denote by $g(r)$ the number of integer points inside the disk $B(0,r)$. We assume that $N_A g(r) \ll N^2$. Define the following subsets of configurations:

- *Self-excluding configurations:* let $\mathscr{E}_{r,N_A,N}$ denote the subset of all point configurations on the square lattice such that no two points are at distance lesser than r;
- *Clustered configurations:* let $\varepsilon \in (0,1)$, define $N_{A,\varepsilon} = \lfloor \varepsilon N_A \rfloor$, and assume further that $N_{A,\varepsilon} g(r) \gg N_A$. Define $\mathscr{C}(\varepsilon, r, N_A, N)$ as the subset of all point configurations $\{x_i,\ i = 1,\ldots,N_A\}$ constructed in the following way:
  - the configuration $\left\{x_1,\ldots,x_{N_{A,\varepsilon}}\right\}$ belongs to $\mathscr{E}(r, N_{A,\varepsilon}, N-r)$ (it is called the *mother configuration*) and $\left\{x_{N_{A,\varepsilon}+1},\ldots,x_{N_A}\right\}$ is a pure random configuration in the sublattice $\bigcup_{i=1}^{N_{A,\varepsilon}} B(x_i,r) \cap \mathbb{Z}^2$ (the *progeny configuration*).

It is clear that if ε and N are chosen correctly those two subsets of configurations are non empty (large), thus one may consider the uniform distribution over those two subsets, giving two random configurations. The generation of such random configurations may be easily performed using the Metropolis-Hastings algorithm[1].

We have performed this simulation on this regular lattice with $N = 250$ and small configurations, and performed a large number of simulations giving similar results. This choice was dictated by the simulation time of the configurations, the parameters chosen giving the desired clustered or diluted behaviors. The confidence intervals were computed using the exact formulas for the variances, with a confidence level of $1 - \alpha = 0.95$.

The numerical data of 30 locations in the square lattice of size 501×501 are simulated with an exclusion radius of 45 for the self-excluding configuration, and the clustered region is made of five clusters of radius 20.

Results are summarized in figure 4: the top picture shows in white diamonds the *intra*-coefficient for a (random) clustered configuration generated by the procedure described above for r ranging from 30 to 70, one observes that the obtained coefficient is always greater than 1. The confidence region is situated under the blue line. The bottom picture represents in black diamonds the *intra*-coefficient for a (random) self-excluding configuration

[1] A fixed configuration belonging to $\mathscr{E}_{r,N_A,N}$ is iteratively modified by moving at random one of its points, the movement is accepted if the new configuration is still in $\mathscr{E}_{r,N_A,N}$. This procedure is repeted a large number of times (see [Häg02]). The clustered configuration is generated this way for the mother configuration, the progeny is generated by an urn without replacement scheme.

generated by the procedure described above for r ranging from 30 to 70, one observes that the obtained coefficient is always lesser than 1. The confidence region is situated above the blue line (note that the two scales are different).
We observe, except for small r for the self-excluding configuration, that the *intra*-coefficient obtained do not belong to the confidence region, this shows that the *intra*-coefficient does actually discriminate clustered, self-excluding and purely random configurations.

Fig. 4 Plot of the *intra*-coefficient for clustered (top) and self-excluding (bottom) configurations.

We may remark that the confidence interval is quite large, due to the fact that the size of the configuration is small, however, the deviation is already siginificant.

4 Continuous setting and asymptotic normality

The computations in a discrete setting are often more complicated than in the case of a continuous setting. This is clearly the case when the question of asymptotic normality is addressed. In this section, we first give the equivalent definitions for the *inter* and *intra*-coefficients in a general continuous setting (allowing spatial heterogeneities) and give the exact computation of their variances. This result is then compared to the discrete one in the large N_t limit. We also prove a central limit theorem for the homogeneous case. This result, although restricted to ideal homogeneous environments, can be interpreted as a justification for the normal approximation for inhomogeneous environments. This approximation is needed to sharpen the confidence interval in the testing procedure (the width of the confidence interval is still proportional to the square root of the variance, but with a smaller constant depending on α and the Normal law).
We begin with a short paragraph on Poisson point processes that will be the natural way to model the *pure random hypothesis* in a continuous framework.

4.1 Poisson point processes

The Poisson point process [Kin93] is the natural way to generate random sets of points in (possibly) unbounded domains. We recall shortly its definition below in a subset of the plane in a diffuse context: let D be a Borel subset of $\mathbb{R}^2$, Λ a diffuse[2] non-negative σ-finite Borel measure on D, the Poisson point process with intensity measure Λ is the random locally finite[3] set of points $\mathbf{X}$ of D such that

- for each Borel subset A of D with finite Λ measure, the number of points of $\mathbf{X}$ in A, denoted by $\sharp(\mathbf{X}\cap A)$, is a Poisson random variable with parameter $\Lambda(A)$:
$$\forall k \geq 0,\ P(\sharp(\mathbf{X}\cap A)=k)=\frac{\Lambda(A)^k}{k!}\exp(-\Lambda(A)),$$
- for disjoint Borel subsets $B_1,\ldots,B_n$ of D, the random variables $(\sharp(\mathbf{X}\cap B_i))$ for $i \in \{1,\ldots,n\}$ are independent.

The term intensity for Λ may be interpreted as *local density*, indeed the average number of points in a bounded subset A is equal to $\Lambda(A)$: as Λ is assumed to be diffuse, there exists a Borel measurable function λ defined on D such that $\Lambda(A)=\int_A \lambda(x)\,dx$, the function λ clearly plays the role of a density.
The Poisson point process has two important properties [Kin93]:

Lemma 4 *Let $\mathbf{X}$ be a Poisson point process with intensity measure Λ on D, and A a bounded Borel subset of D, then, conditionnally on the event $\{\sharp(\mathbf{X}\cap A)=k\}$ where $k \geq 1$, the points of $\mathbf{X}\cap A$ are distributed as k independent points with common law $\Lambda/\Lambda(A)$.*

This property is also a characteristic of the Poisson point process. The important other property is the Campbell-Mecke-Slivnyak formula stating that the Poisson point process conditionned on the negligible event that $\sharp(\mathbf{X}\cap\{x\})=1$ is equal in law to a Poisson point process with the same intensity measure *plus* the fixed point x:

Lemma 5 *Let f be a non negative function defined on the product space $D\times\mathscr{P}_{l.f.}(D)$, where $\mathscr{P}_{l.f.}$ is the set of all locally finite subsets of D, and let $D'\subset D$, then one has for a Poisson Point Process $\mathbf{X}$ with intensity measure Λ on D:*

$$E\left[\sum_{x\in\mathbf{X}\cap D'} f(x,\mathbf{X})\right]=\int_{D'} E\left[f(x,\mathbf{X}\cup\{x\})\right]\,d\Lambda(x).$$

In this formula, a sum indexed by an empty set is assumed to be zero.

In the following sections, the locations of stores of types A and will be the points of a Poisson point process $\mathbf{X}_A$ with given intensity Λ_A when we are dealing with the *intra*-coefficient, and we shall write $N_A(x,r)=\sharp(\mathbf{X}_A\cap B(x,r))$.
When dealing with the *inter*-coefficient, the locations of the points of type A are assumed to be fixed, and the B stores are the points of a Poisson point process $\mathbf{X}_B$ with given intensity Λ_B, and we denote $N_B(x,r)=\sharp(\mathbf{X}_B\cap B(x,r))$.
The notation $N_t(x,r)$ has no equivalent in this continuous settting.

Both intensities Λ_A and Λ_B are supposed to be *finite* ($\Lambda_A(D),\Lambda_B(D)<+\infty$), thus the total number of A and B stores is almost surely finite.

[2] By diffuse we mean absolutely continuous with respect to the Lebesgue measure.

[3] By locally finite we mean that for each compact subset K of D, almost surely the number of points of $\mathbf{X}$ in K is finite.

4.2 Definition of the *inter* and *intra*-coefficients

The generalization of the a_{AB} coefficient is obtained by considering the relative weights of some neighborhoods for the Λ_B measure, indeed rewrite the discrete a_{AB} in the following way:

$$a_{AB} = \frac{1}{N_A} \sum_{i=1}^{N_A} \frac{N_B(A_i,r)}{N_B} \left(\frac{N_t(A_i,r) - N_A(A_i,r)}{N_t - N_A} \right)^{-1},$$

one observes that the right-most term is the inverse of the relative weight of the neighborhood $B(A_i,r)$ in the set T. The analog of this formula is thus given by

$$\begin{aligned} a^P_{AB} &= 1 \text{ if } N_A = 0, \\ &= \frac{1}{N_A} \sum_{i=1}^{N_A} \frac{N_B(A_i,r)}{N_B} \left(\frac{\Lambda_B(V_i)}{\Lambda_B(D)} \right)^{-1}, \text{ otherwise,} \end{aligned}$$

where $V_i = B(A_i,r) \cap D$.

The computation of the average (expectation) of this coefficient is straightforward using the definition of the Poisson point process $\mathbf{X}_B$, one checks easily that it is equal to 1.

If one rewrites the discrete a_{AA} in the following way:

$$a_{AA} = \frac{1}{N_A} \sum_{i=1}^{N_A} \frac{N_A(A_i,r)}{N_A - 1} \left(\frac{N_t(A_i,r)}{N_t - 1} \right)^{-1}.$$

The natural extension of the a_{AA} term for the Poisson case becomes in the same way as above:

$$\begin{aligned} a^P_{AA} &= 1 \text{ if } N_A = 0, \\ &= \frac{1}{N_A} \sum_{i=1}^{N_A} \frac{N_A(A_i,r)}{N_A - 1} \left(\frac{\Lambda_A(V_i)}{\Lambda_A(D)} \right)^{-1} \text{ otherwise.} \end{aligned}$$

Using lemma 5 it is clear that the average value of this *intra*-coefficient is also equal to 1.

4.3 Computation of the variances

In order to compute the variance of the *inter* and *intra*-coefficients, we need an extended version of lemma 5 ([Kin93]):

Lemma 6 *Let f be a non negative function defined on the product space $D^2 \times \mathscr{P}_{l.f.}(D)$, and let $D' \subset D$, then one has for a Poisson Point Process $\mathbf{X}$ with intensity measure Λ on D:*

$$E\left[\sum_{x \neq y \in \mathbf{X} \cap D'} f(x,y,\mathbf{X}) \right] = \int_{D'^2} E\left[f(x,y,\mathbf{X} \cup \{x,y\}) \right] d\Lambda(x)\, d\Lambda(y).$$

This lemma is sometimes stated with symmetric functions in their first two arguments $g(x,y,\mathbf{X}) = g(y,x,\mathbf{X})$ and sums over pairs of distinct points:

$$E\left[\sum_{\{x,y\} \subset \mathbf{X} \cap D'} g(x,y,\mathbf{X}) \right] = \frac{1}{2} \int_{D'^2} E\left[g(x,y,\mathbf{X} \cup \{x,y\}) \right] d\Lambda(x)\, d\Lambda(y).$$

Using this lemma the computation of the variances is straightforward (though a little tricky), and we get

Proposition 4 *The variance of the* inter-*coefficient is given by*

$$\mathrm{Var}(a^P_{AB}) = \left(-1+\frac{1}{N_A}\left\langle\frac{1}{p^B_i}\right\rangle_A+\frac{N_A-1}{N_A}\left\langle\frac{p^B_{ij}}{p^B_i p^B_j}\right\rangle_A\right)E[N_B^{-1};N_B>0],$$

where the quantities p^B_i and p^B_{ij} are given by

$$p^B_i=\frac{\Lambda_B(V_i)}{\Lambda_B(D)},\ p^B_{ij}=\frac{\Lambda_B(V_i\cap V_j)}{\Lambda_B(D)},$$

and the averages $\langle\cdot\rangle_A$ are taken with respect to the fixed points of type A as in the first sections, and the last term is equal to

$$E[N_B^{-1};N_B>0]=\sum_{k\geq 1}\frac{\Lambda_B(D)^k}{k k!}\exp(-\Lambda_B(D)).$$

The computation of the *intra* variance is a bit more complicated, using the same arguments as in proposition 4 we obtain

Proposition 5 *The variance of the* intra-*coefficient is the sum of the following terms:*

$$\begin{aligned}\mathrm{Var}(a^P_{AA}) = &-\Lambda_A(D)^2 E\left[\frac{1}{(N_A+1)(N_A+2)^2}\right]\\ &-\Lambda_A(D)E\left[\frac{1}{N_A(N_A+1)^2};N_A>0\right]\\ &+E\left[\frac{1}{N_A(N_A+1)^2};N_A>0\right]\int_D\frac{1}{p(a)}\,d\Lambda_A(a)\\ &+E\left[\frac{N_A}{(N_A+1)^2(N_A+2)^2}\right]\int_{D^2}\frac{p(a,b)}{p(a)p(b)}\,d\Lambda_A(a)\,d\Lambda_A(b)\\ &+E\left[\frac{1}{(N_A+1)^2(N_A+2)^2}\right]\int_{D^2}\frac{\mathbf{1}_{a\sim b}}{p(a)p(b)}\,d\Lambda_A(a)\,d\Lambda_A(b),\end{aligned}$$

where $p(a)=\Lambda_A(B(a,r)\cap D)/\Lambda_A(D)$, $p(a,b)=\Lambda_A(B(a,r)\cap B(b,r)\cap D)/\Lambda_A(D)$, $\mathbf{1}_{a\sim b}=1$ if $|a-b|<r$, 0 otherwise, and

$$E[Z(N_A);N_A>0]=\sum_{n=1}^{+\infty}Z(n)\,\frac{\Lambda_A(D)^n}{n!}\exp(-\Lambda_A(D)).$$

4.4 Comparison with the discrete case: asymptotic equivalence

Classically the Poisson (point) process can be seen as a limit of some discrete model, thus we therefore expect to obtain similar values for both approaches in large domains (equivalently N_t large). Let us start with the discrete case: For $N_t \gg 1$, $N_A \gg 1$, $N_t \gg N_A$, $N_t \gg N_B$, the variance of a_{AB} becomes:

$$\sigma^2(a_{AB})\simeq\frac{1}{N_B}\left(-1+\frac{N_t}{N_A}\left\langle\frac{1}{N_t(A_i)-N_A(A_i)}\right\rangle_A+N_t\left\langle x_{ij}\right\rangle_A\right),$$

which would be indistinguishable in Figure 1 from the complete theoretical expression. It appears that $\mathrm{Var}(a_{AB})$ decreases roughly as $1/N_B$ for fixed A's. The last two terms, which

characterize the distribution of A, can be more or less important according to the spatial distribution of A stores.
Under the same conditions, a_{AA} variance can be approximated by:

$$\sigma^2(a_{AA}) \simeq \frac{1}{N_A}\left(-1 + N_t \left\langle x_{ij}\right\rangle_t + \frac{N_t}{N_A}\left\langle \frac{1}{N_t^i}\right\rangle_t + \frac{N_t{}^2}{N_A}\left\langle \frac{1}{N_t^i N_t^j}\right\rangle_n\right),$$

which, again, is indistinguishable in figure 2 from the complete theoretical expression[4], and $\mathrm{Var}(a_{AA})$ decreases roughly as $1/N_A$.

To compare these limits with the continuous case, let us display the results side by side:

Discrete setting	Continous (Poisson) setting
$\sigma^2(a_{AB}) = -\dfrac{1}{N_B} +$	$\sigma^2(a^P_{AB}) = -E[N_B^{-1}; N_B > 0] +$
$\dfrac{N_t}{N_A N_B}\left\langle \dfrac{1}{N_t(A_i) - N_A(A_i)}\right\rangle_A$	$\dfrac{1}{N_A} E[N_B^{-1}; N_B > 0]\left\langle \dfrac{1}{p_i^B}\right\rangle_A +$
$\dfrac{N_t}{N_B}\left\langle x_{ij}\right\rangle_A,$	$\dfrac{N_A - 1}{N_A} E[N_B^{-1}; N_B > 0]\left\langle \dfrac{p_{ij}^B}{p_i^B p_j^B}\right\rangle_A.$

Table 1 Comparison of the expressions of the discrete and continuous variances of the *inter*-coefficient.

The first lines are clearly similar. To see the similarity of the last two lines, one has to replace the quantities $N_t\left\langle 1/(N_t(A_i) - N_A(A_i))\right\rangle_A$ and $N_t\left\langle x_{ij}\right\rangle_A$ by their respective values

$$\begin{aligned} N_t\left\langle \frac{1}{N_t(A_i) - N_A(A_i)}\right\rangle_A &= \left\langle \frac{1}{(N_t(A_i) - N_A(A_i))/N_t}\right\rangle_A \\ &\sim \left\langle \frac{1}{p_i^B}\right\rangle_A, \end{aligned}$$

$$\begin{aligned} N_t\left\langle x_{ij}\right\rangle_A &= \left\langle \frac{(N_t(C_{i,j}) - N_A(C_{i,j}))/N_t}{((N_t(A_i,r) - N_A(A_i,r))/N_t)((N_t(A_j,r) - N_A(A_j,r))/N_t}\right\rangle_A, \\ &\sim \left\langle \frac{p_{ij}^B}{p_i^B p_j^B}\right\rangle_A. \end{aligned}$$

For the *intra* terms, the comparison proceeds in the same manner, in the homogenous case with intensity measure Λ_A a multiple of the Lebesgue measure: $d\Lambda_A = \lambda_A\, dx$, define

$$\overline{N_A} = \lambda_A |D|, \text{ and } \overline{n_r} = \lambda_A \pi r^2$$

[4] For Lyon's configuration where $N_t\left\langle x_{ij}\right\rangle_t = 0.933$, $N_t\left\langle 1/N_t^i\right\rangle_t = 1040$ and $N_t{}^2\left\langle 1/(N_t^i N_t^j)\right\rangle_n = 793$, the two last terms dominate as long as $N_A \ll 300$, while for $N_A \gg 300$ all terms become of the same order of magnitude.

the mean number of shops in the whole domain and in a neighborhood of radius r, we obtain the following asymptotics for large domains

$$\sigma^2(a^P_{AA}) \simeq \frac{2}{\overline{N_A}\,\overline{n_r}}.$$

From this asymptotic relation we deduce that $\sqrt{\overline{N_A}}(a^P_{AA} - 1)$ has a finite variance in the limit as the domain D increases: in the next subsection we prove in the homogeneous case an asymptotic normality property that explains and extends this remark.

4.5 Asymptotic normality in the homogeneous case

The asymptotic normality of coefficients such as the one studied here does not seem to have been studied in the litterature, as said in the introduction this property gives sharpened confidence intervals for the testing of the *pure randomness* hypothesis. We carried out some tests with Monte Carlo simulations on actual subsets of sites of the city of Lyon:

- For the *inter*-coefficient, we have chosen to simulate 87 B sites around 917 A sites that are strongly aggregated (*intra*-coefficient $a_{AA} = 2.17$). The 87 B sites are chosen randomly among the 6922 free T sites. The circles correspond to the evaluation of the probability density function of a_{AB} at different points in $[0.6, 1.4]$ performed by direct simulation. Note that the distribution in this A configuration is more subject to large fluctuations due to the strong aggregation. The result in Figure 5 is compared to the Normal law centered at 1 with variance computed from proposition 1.
- In the a_{AA} case, the distribution is also indistinguishable from the Normal law, as shown in Figure 6. This figure is obtained by randomly choosing 917 A sites among 7839 T sites. The continuous line in figure 6 corresponds to a Normal law centered at 1 with variance computed from proposition 2.

The adequation of the estimated distributions with the corresponding Normal laws seems to be an indicator that such asymptotic normality actually exists.

In the following, we shall prove that these distributions do converge to Normal laws when the domain (or the intensity) is large in the *intra* case, with *constant intensity*:

$$d\Lambda_A = \lambda_A\, dx.$$

Fig. 5 Distribution of random a_{AB} values compared to the Normal law centered at 1 with variance computed from proposition 1.

Fig. 6 Distribution of random a_{AA} values compared to a Normal law centered at 1 with variance computed from proposition 2.

Let us rewrite a^P_{AA} on the square domain $D^n = [0,n)^2 = \bigcup_{i,j=0}^{n-1}[i,i+1) \times [j,j+1)$ (those smaller squares being denoted by $D_{i,j}$):

$$\begin{aligned}
a^P_{AA}(D^n) &= \sum_{a \in \mathbf{X}_A \cap D^n} \frac{1}{N_A(D^n) p_n(a)} \left(\frac{N_A(a)}{N_A(D^n) - 1} \right) \\
&= \sum_{i,j=0}^{n-1} \left(\sum_{a \in \mathbf{X}_A \cap D_{i,j}} \frac{1}{N_A(D^n) p_n(a)} \left(\frac{N_A(a)}{N_A(D^n) - 1} \right) \right), \\
&= \sum_{i,j=0}^{n-1} \left(\sum_{a \in \mathbf{X}_A \cap D_{i,j}} \frac{\lambda_A(D^n)}{N_A(D^n)} \frac{1}{\lambda_A(V(a))} \left(\frac{N_A(a)}{N_A(D^n) - 1} \right) \right), \\
&= \underbrace{\frac{\lambda_A(D^n)}{N_A(D^n)}}_{\xrightarrow[n \to +\infty]{\text{a.s.}} 1} \underbrace{\frac{n^2}{N_A(D^n) - 1}}_{\xrightarrow[n \to +\infty]{\text{a.s.}} \lambda_A^{-1}} \left\{ \frac{1}{n^2} \sum_{i,j=0}^{n-1} \underbrace{\left(\sum_{a \in \mathbf{X}_A \cap D_{i,j}} \frac{N_A(a)}{\lambda_A(V(a))} \right)}_{Y_{i,j}} \right\}.
\end{aligned}$$

It is straightforward to adapt the classical mixing central limit theorem for sequences to a two-dimensional setting: the random variables $(Y_{i,j})_{i,j \geq 0}$ form a mixing sequence (as they are indeed independent at large distance), so that we easily get

Theorem 1 *As n tends to infinity one has*

$$\sqrt{N_A(D^n)} \left(a^P_{AA} - \frac{\lambda_A^2 n^4}{N_A(D^n)(N_A(D^n) - 1)} \right) \xrightarrow[n \to +\infty]{law} \mathcal{N}\left(0, 4 + \frac{2}{\lambda_A \pi r^2}\right).$$

This theorem has a bias in the mean, and therefore in the asymptotic variance. This may be corrected as the *intra* term may also be written as a V-statistic:

$$a^P_{AA} = \frac{\lambda_A(D)}{N_A(N_A - 1)} \sum_{x \neq y \in \mathbf{X}_A} \frac{\mathbf{1}_{|x-y| \leq r}}{\lambda_A(V(x))}.$$

For sake of simplicity we shall still work in the square D_n. Let $(Y_k)_{k \geq 1}$ be independent identically distributed random variables, uniformely distributed on $[0,1)^2$ and N_n an independent Poisson random variable with parameter $\lambda_A n^2$, so that

$$\begin{aligned}
a^P_{AA}(D^n) - 1 &= \frac{\lambda_A n^2}{N_n(N_n - 1)} \sum_{1 \leq i \neq j \leq N_n} \frac{\mathbf{1}_{n|Y_i - Y_j| \leq r}}{\lambda_A(B(nY_i, r))} - 1, \\
&= \frac{1}{N_n(N_n - 1)} \sum_{1 \leq i \neq j \leq N_n} \left(\frac{\mathbf{1}_{|Y_i - Y_j| \leq r/n}}{|B(Y_i, r/n) \cap (0,1)^2|} - 1 \right),
\end{aligned}$$

this is a sort of V-statistic, indexed by the random variable N_n, where the summand will be denoted by $G^n_{i,j}$ (it is not symmetric). This random variable

- is centered,
- bounded by $1 + 4n^2/(\pi r^2)$,

- and satisfies, denoting $p_{r/n}(Y) = |B(Y,r/n) \cap [0,1)^2|$ and $p_{r/n}(Y_1,Y_2) = |B(Y_1,r/n) \cap B(Y_2,r/n) \cap [0,1)^2|$,

$$
\begin{aligned}
E[(G^{(n)}_{1,2})^2] &= E\left[\frac{1}{p_{r/n}(Y)}\right] - 1,\\
E[G^{(n)}_{1,2}G^{(n)}_{2,1}] &= E\left[\frac{\mathbf{1}_{|Y_1-Y_2|\le r/n}}{p_{r/n}(Y_1)p_{r/n}(Y_2)}\right] - 1,\\
E[G^{(n)}_{1,2}G^{(n)}_{2,3}] &= 0,\\
E[G^{(n)}_{1,2}G^{(n)}_{1,3}] &= 0,\\
E[G^{(n)}_{1,2}G^{(n)}_{3,2}] &= E\left[\frac{p_{r/n}(Y_1,Y_2)}{p_{r/n}(Y_1)p_{r/n}(Y_2)}\right] - 1,\\
E[G^{(n)}_{1,2}G^{(n)}_{3,1}] &= 0,\\
E[G^{(n)}_{1,2}G^{(n)}_{3,4}] &= 0.
\end{aligned}
$$

From those relations we have the following asymptotics

$$
\begin{aligned}
E[(G^{(n)}_{1,2})^2] &\sim \frac{n^2}{\pi r^2},\\
E[G^{(n)}_{1,2}G^{(n)}_{2,1}] &\sim \frac{n^2}{\pi r^2},\\
E[G^{(n)}_{1,2}G^{(n)}_{3,2}] &\le Cn^{-2}.
\end{aligned}
$$

This asymptotic degeneracy of most of the terms in the moments above changes the usual speed of the central limit theorem for V-statistics, and gives the following central limit result:

Theorem 2 *As the domain D tends towards* $\mathbb{R}^2$ *in a* regular *way, then*

$$
\sqrt{N_A(D)}\,\left(a^P_{AA} - 1\right) \xrightarrow[D\to\mathbb{R}^2]{law} \mathscr{N}\left(0, \frac{2}{\lambda_A \pi r^2}\right).
$$

The proof of this result is rather technical, it proceeds by the computation of the moments of the left hand side and of their asymptotics. An extended sketch of the proof is given in the appendix.

In this result we see once again the right asymptotics for the variance, $\sigma^2(a^P_{AA}) \sim 2/(\lambda_A \pi r^2 \lambda_A |D|)$.

The existence of central limit theorems for inhomogeneous intensities and/or for the *inter*-coefficients is still an open problem in this setting, even if numerical evidence seems to justify this approximation.

5 Differential coefficients and comparison with the Duranton-Overman indicators

In this section we shortly discuss the diferentiated version of the *inter* and *intra*-coefficients introduced above, and eventually show some comparative results with the indicator introduced in [DO05].

5.1 Differential coefficients

The differentiated version of the *inter* and *intra*-coefficients is the discrete derivation of a_{AB} and a_{AA}. Let δr be positive, and define $N_A(A_i, r, r+\delta r)$ as the number of points of type A in the shell $r \leq |A_i - x| < r + \delta r$, let us give the following notations:

$$\begin{aligned} N_A(A_i, r, r+\delta r) &= \text{number of } A \text{ points } x \text{ such that } r \leq |A_i - x| < r+\delta r, \\ N_B(A_i, r, r+\delta r) &= \text{number of } B \text{ points } x \text{ such that } r \leq |A_i - x| < r+\delta r, \\ \tilde{N}_t^i(r, \delta r) &= \text{number of points } x \text{ of } T \text{ such that } r \leq |T_i - x| < r+\delta r. \end{aligned}$$

Then the formulas giving the variance for the differentiated indicator are exactly the same as before, using those new quantities. The indicators are:

$$d_{AB} := \frac{N_t - N_A}{N_A N_B} \sum_{i=1}^{N_A} \frac{N_B(A_i, r, r+\delta r)}{N_t(A_i, r, r+\delta r) - N_A(A_i, r, r+\delta r)} \tag{11}$$

$$d_{AA} := \frac{N_t - 1}{N_A(N_A - 1)} \sum_{i=1}^{N_A} \frac{N_A(A_i, r, r+\delta r)}{N_t(A_i, r, r+\delta r)} \tag{12}$$

and

Proposition 6 *The variance of the* Differential inter-*coefficient is given by*

$$\begin{aligned} \sigma^2(d_{AB}) = &\left(\frac{N_t - N_A}{N_A N_B} \frac{N_t - N_A - N_B}{N_t - N_A - 1}\right) \left\langle \frac{1}{N_t(A_i, r, r+\delta r) - N_A(A_i, r, r+\delta r)} \right\rangle_A \\ &- \frac{N_t - N_A - N_B}{N_B(N_t - N_A - 1)} \\ &+ \langle y_{ij} \rangle_A \left(\frac{N_t - N_A - N_B}{N_t - N_A - 1}\right) \left(\frac{N_t - N_A}{N_B}\right) \left(1 - \frac{1}{N_A}\right), \end{aligned}$$

where

$$\langle y_{ij} \rangle_A = \left\langle \frac{N_t(C'_{i,j}) - N_A(C'_{i,j})}{(N_t(A_i, r, r+\delta r) - N_A(A_i, r, r+\delta r))(N_t(A_j, r, r+\delta r) - N_A(A_j, r, r+\delta r)} \right\rangle_A,$$

and $C'_{i,j} = \{x \in \mathbb{R}^2 \ : \ r \leq |x - A_i| < r + \delta r \textit{ and } r \leq |x - A_i| < r + \delta r\}$.
The variance of the Differential intra-*coefficient is the sum of four terms:*

$$\sigma^2(d_{AA}) = \sum_{i=1}^{4} \mathrm{Var}(d_{AA})_i,$$

where

$$\begin{aligned} \mathrm{Var}(d_{AA})_1 &= \frac{N_t - N_A}{(N_t - 2)(N_A - 1)}, \\ \mathrm{Var}(d_{AA})_2 &= \frac{(N_t - 1)(N_t - N_A)}{N_A(N_A - 1)(N_t - 2)} \left\langle \frac{1}{\tilde{N}_t^i(r, \delta r)} \right\rangle_t, \\ \mathrm{Var}(d_{AA})_3 &= \frac{(N_t - 1)^2(N_t - N_A)(N_t - N_A - 1)}{(N_t - 2)(N_t - 3)N_A(N_A - 1)} \left\langle \frac{1}{\tilde{N}_t^i(r, \delta r)\tilde{N}_t^j(r, \delta r)} \right\rangle_n, \\ \mathrm{Var}(d_{AA})_4 &= \frac{(N_t - 1)^2(N_t - N_A)(N_A - 2)}{(N_t - 2)(N_t - 3)N_A(N_A - 1)} \langle z_{ij} \rangle_t, \end{aligned}$$

where $\langle z_{ij} \rangle_t$ *is defined analogously to* $\langle y_{ij} \rangle_A$*:*

$$\langle z_{ij} \rangle_t = \frac{2}{N_t(N_t-1)} \sum_{1 \le i < j \le N_t} \frac{N_t(\tilde{T}^{r,\delta r}_{i,j})}{N_t(T_i, r, r+\delta r) N_t(T_j, r, r+\delta r)},$$

where $\tilde{T}^{r,\delta r}_{i,j} = \{x \in \mathbb{R}^2 \ : \ r \le |x - T_i| < r + \delta r$ *and* $r \le |x - T_i| < r + \delta r\}$.

5.2 Duranton-Overman indicators

This differential coefficient is very close to the indicator introduced by Duranton and Overman, let us recall their definition: let f be a non negative kernel with total mass 1, $h > 0$ be the typical length scale of the discretisation, then $\widehat{K}$ is defined as the kernel density estimate for n points with inter-distances $d_{i,j}$

$$\widehat{K}_{f,h}(r) = \frac{2}{hn(n-1)} \sum_{1 \le i < j \le n} f\left(\frac{r - d_{i,j}}{h}\right),$$

if we rewrite this with the particular kernel $f_0(x) = \mathbf{1}_{(-1,0]}(x)$ and $h = \delta r$, this gives for N_A points of type A

$$\widehat{K}_{f_0,\delta r}(r) = \frac{1}{N_A(N_A-1)} \sum_{i=1}^{N_A} \frac{N_A(A_i, r, r+\delta r)}{\delta r}.$$

The computation of the expectation of this coefficient is very similar to the computations detailed before

Lemma 7 *For all* $r > 0$ *and* $\delta r > 0$*, we have under the* pure randomness hypothesis*:*

$$E\left[\widehat{K}_{f_0,\delta r}(r)\right] = \frac{\langle \tilde{N}^i_t(r,\delta r) \rangle_t}{\delta r (N_t - 1)}.$$

The computation of the variance is still a little more tricky, but we obtain after a few steps of calculations:

Proposition 7 *For all* $r > 0$ *and* $\delta r > 0$*, we have under the* pure randomness hypothesis*:*

$$\begin{aligned}
\sigma^2\left(\widehat{K}_{f_0,\delta r}(r)\right) = \ & \frac{1}{(\delta r)^2}\left\{\frac{1}{N_A(N_A-1)}\left[\frac{N_A-2}{(N_t-1)(N_t-2)}\frac{2N_t-N_A-3}{N_t-3}\langle \tilde{N}^i_t(r,\delta r)^2 \rangle_t + \right.\right. \\
& \left(1 + \frac{N_t - N_A}{(N_t-1)(N_t-2)} + \frac{(N_A-2)(N_A-3)}{(N_t-2)(N_t-3)}\right)\langle \tilde{N}^i_t(r,\delta r) \rangle_t + \\
& \frac{(N_A-2)(N_A-3)}{(N_t-2)(N_t-3)}\left\langle \tilde{N}^i_t(r,\delta r)\tilde{N}^j_t(r,\delta r) \right\rangle_t + \\
& \left.\left.\frac{(N_A-2)(N_t-N_A)}{(N_t-2)(N_t-3)}\left\langle \tilde{N}^{(ij)}_t \right\rangle_t\right] - \frac{1}{(N_t-1)^2}\langle \tilde{N}^i_t(r,\delta r) \rangle_t^2\right\},
\end{aligned}$$

where $\left\langle \tilde{N}^{(ij)}_t \right\rangle_t$ is the following average over all couples of points of T:

$$\left\langle \tilde{N}^{(ij)}_t \right\rangle_t = \frac{2}{N_t(N_t-1)} \sum_{1 \le i < j \le N_t} N_t(\tilde{T}^{r,\delta r}_{i,j}).$$

The computation of the variance for general kernels f may also be performed along the same lines.

5.3 Drawbacks and advantages of the different coefficients

The two coefficients proposed by Duranton and Overman [DO05] and Marcon and Puech [MP07] share a number of advantages detailed in the introduction (inhomogeneous underlying space used as reference, precise location data (x and y coordinates)). Their main difference is the integrative or differential approach they use. Duranton and Overman focus on the distribution between two distances r and $r+\delta r$, while Marcon and Puech integrate the distribution from 0 to r. The differential coefficient allows to zoom on precise distances, and measure differences from randomness in more detail. In contrast, our coefficient, inspired from Marcon and Puech's approach, is simpler to interpret because the coefficient converges to 1 as r approaches the system size, thus allowing to readily quantify deviations from randomness. In Duranton and Overman's approach, absolute values of the coefficients are meaningless. It is the deviations to the random values which show whether the spatial distribution is aggregated or dispersed.

6 Conclusions and perspective

This paper gives analytic formulas to compute the variance of some coefficients of purely random distributions of points (meaning non interacting distributions). We *rigorously* show that these distributions asymptotically follow a Normal law. Our paper allows to dispense with Monte Carlo simulations, which can be cumbersome to implement and prohibitively time consuming for large samples. Our analytical expressions may also allow to understand qualitatively the main factors that may change the variance: number of sites, spatial inhomogeneities, etc.

A natural extension is to get a better mathematical understanding of a given situation, whether clustered or excluding, by a precise description of the (random) point process having the same behavior. The first way to introduce non independence between sites is to consider Gibbs point processes (or Markov point processes), characterized by an interaction potential, as well as a general potential linked with the *landscape*, taking into account for instance the population density, or some other geographical or economical artefact. The main topics in this framework shall be the estimation of the actual characteristics (or parameters) of those potentials, and the comparison of those parameters for different economic/geographic situations. This work is in progress.

References

[BCL07] A. Briant, P.-P. Combes, and M. Lafourcade. Do the size and shape of spatial units jeopardize economic geography estimations? http://www.vcharite.univ-mrs.fr/PP/combes/maup.pdf, (2007).

[Bes77] J. E. Besag. Comments on Ripley's paper. *Journal of the Royal Statistical Society B*, 39:193–195, (1977).

[Col06] Collective. Competition complementarity in retailing. *Revue Belge de géographie*, (1-2), (2006).

[DO05] G. Duranton and H. G. Overman. Testing for localisation using micro-geographic data. *The Review of Economic Studies*, 72:1077, (2005).

[EB03] T. Egami and S. Billinge. *Underneath the Bragg Peaks: strutural analysis of complex materials*. Material Series. Pergamon, (2003).

[EG97] G. Ellison and E. L. Glaeser. Geographic concentration in us manufacturing industries: A dartboard approach. *Journal of Political Economy*, 105(5):889–927, (1997).

[Häg02] O. Häggström. *Finite Markov chains and algorithmic applications*, vol. 52 of London Mathematical Society Student Texts. Cambridge University Press, Cambridge, (2002).

[HG84] E. M. Hoover and F. Giarratani. An introduction to regional economics. http://www.rri.wvu.edu/WebBook/Giarratani/contents.htm, (1984).

[Hoo37] E. M. Hoover. *Location theory and the shoe and leather industries*. Cambridge, MA: Harvard University Press, (1937).

[Jen06] Pablo Jensen. Network-based predictions of retail store commercial categories and optimal locations. *Phys. Rev. E*, 74, (2006).

[Kin93] J. F. C. Kingman. *Poisson processes*, volume 3 of *Oxford Studies in Probability*. The Clarendon Press Oxford University Press, New York, (1993).

[MNPT05] Y. Manolopoulos, A. Nanopoulos, A. N. Papadopoulos, and Y. Theodoridis. *R-trees: Theory and Applications*. Springer-Verlag, (2005).

[MP07] E. Marcon and F. Puech. Measures of the geographic concentration of industries: Improving distance-based methods. `http://halshs.archives-ouvertes.fr/halshs-00372617/fr/`, (2009).

[MS99] F. Maurel and B. Sedillot. A measure of the geographic concentration of french manufacturing industries. *Regional Science and Urban Economics*, 29(5):575–604, (1999).

[Ope84] S. Openshaw. *The Modifiable Areal Unit Problem*. Norwich: Geo Books, (1984).

[Par98] Katy Paroux. Quelques théorèmes centraux limites pour les processus Poissoniens de droites dans le plan. *Adv. in Appl. Probab.*, 30(3):640–656, (1998).

[Rip76] B. D. Ripley. The second-order analysis of stationary point processes. *J. Appl. Probability*, 13(2):255–266, (1976).

[Unw96] D. J. Unwin. Gis, spatial analysis and spatial statistics. *Progress in Human Geography*, 20:540–551, (1996).

[WPF96] J. S. Ward, G. R. Parker, and F. J. Ferrandino. Long-term spatial dynamics in an old-growth deciduous forest. *For. Ecol. Manage.*, 83:189–202, (1996).

[YK50] G. Udny Yule and M. G. Kendall. *An Introduction to the Theory of Statistics*. Hafner Publishing Co., New York, N. Y., (1950). 14th ed.

Appendix: proof of theorem 2

Proof We give the proof only the square case, as the generalization to domains D converging to $\mathbb{R}^2$ in a regular way (that is with a bounded ratio between its diameter and the radius of the greatest disc included in D) follows the same lines.
Let us proceed as the for the general classical combinatorial proof of the central limit theorem for V-statistics: this proof starts with the computation of the moments:

$$\begin{aligned} E\left[\left(a^P_{AA}(D^n)-1\right)^k\right] &= \sum_{N\geq 0} E\left[\left(a^P_{AA}(D^n)-1\right)^k \mid N_n = N\right] P(N_n = N), \\ &= \sum_{N\geq 2} \frac{1}{(N(N-1))^k} E\left[\left(\sum_{i,j=1}^{N} G^{(n)}_{i,j}\right)^k\right] \frac{(\lambda_A n^2)^N}{N!} \exp(\lambda_A n^2), \end{aligned}$$

the inner expectation can be expanded as

$$E\left[\sum_{i_1,j_1,\ldots,i_k,j_k=1}^{N} \prod_{s=1}^{k} G^{(n)}_{i_s,j_s}\right] = \sum_{i_1,j_1,\ldots,i_k,j_k=1}^{N} E\left[\prod_{s=1}^{k} G^{(n)}_{i_s,j_s}\right].$$

This product is equal to zero as soon as there exists an index s_0 such that $\{i_{s_0}, j_{s_0}\}$ does not intersect the set $\{i_s, j_s; s \neq s_0\}$. For any choice of the indices, the expectation of such a product is bounded by $(1+4n^2/(\pi r))^k \sim Cn^{2k}$. Our purpose is now to count the terms in this sum that do really matter: from the first remark above, we may introduce the following equivalence relation, let $(\mathbf{i},\mathbf{j}) = \{(i_1, j_1), \dots, (i_k, j_k)\}$, and (i, j) and (k,l) two couples of indices in this set, then

$$\begin{aligned}(i,j) \sim (k,l) \text{ if and only if } &\exists (i_1, j_1), \dots, (i_t, j_t) \in (\mathbf{i},\mathbf{j}) \text{ such that}\\ &\{i, j\} \cap \{i_1, j_1\} \neq \emptyset, \dots, \{i_u, j_u\} \cap \{i_{u+1}, j_{u+1}\} \neq \emptyset, \dots,\\ &\text{and } \{i_t, j_t\} \cap \{k,l\} \neq \emptyset.\end{aligned}$$

Then $(\mathbf{i},\mathbf{j})$ is the disjoint union of classes for this relation, and as soon as a class is a singleton, the expectation of the product, denoted by $P_{(\mathbf{i},\mathbf{j})}$ is zero. Hence all the classes denoted by $(\mathbf{i},\mathbf{j})_1, \dots (\mathbf{i},\mathbf{j})_v$ must have a cardinality greater or equal than 2 in order to have a non zero term, and $P_{(\mathbf{i},\mathbf{j})}$ is equal to the product of the expectations on each class:

$$\begin{aligned}P_{(\mathbf{i},\mathbf{j})} &= \prod_{w=1}^{v} E\left[\prod_{(i,j)\in(\mathbf{i},\mathbf{j})_w} G_{i,j}^{(n)}\right],\\ &= \prod_{w=1}^{v} P_{(\mathbf{i},\mathbf{j})_w}.\end{aligned}$$

From now on the *type* $t_{(\mathbf{i},\mathbf{j})}$ of $(\mathbf{i},\mathbf{j})$ will be the (ordered) sequence of the cardinalities of its classes.

- Let us first assume that $k = 2p$, if $t_{(\mathbf{i},\mathbf{j})} = (a_1, \dots, a_v)$, the number of degrees of freedom for the choice of such indices $(\mathbf{i},\mathbf{j})$ is at most $N^{\sum_{w=1}^{v}(a_w+1)}$, that is N^{v+2p}. As each class has at least two elements one has $v \leq p$, so that the number of degrees of freedom is at most N^{3p}.
  Let us assume that $(\mathbf{i},\mathbf{j})_w$ is reordered as $\{((i_1, j_1)), \dots, ((i_1, j_1)), \dots\}$ where $((i, j))$ denotes either (i, j) or (j,i). We already know that if this class $(\mathbf{i},\mathbf{j})_w$ has two elements ($a_w = 2$), the value of the corresponding expected product is either:
  - $\sim n^2/(\pi r^2)$ if it is $\{((i, j)), ((i, j))\}$,
  - or Cn^{-2} if it is $\{((i, j)), ((j,k))\}$.

  Hence for $t_{(\mathbf{i},\mathbf{j})} = \{2, \dots, 2\}$, the value of $P_{(\mathbf{i},\mathbf{j})}$ is either
  - $\sim (n^2/(\pi r^2))^p$ if each class is of the form $\{((i, j)), ((i, j))\}$,
  - $Cn^{2(p-2p')}$ if there are $(p - p')$ classes of the form $\{((i, j)), ((i, j))\}$, and p' classes of the form $\{((i, j)), ((j,k))\}$, with $p' \in \{1, \dots, p\}$.

  The number of terms of the first type is N^{2p}, whereas the number of the second type of terms is of order $N^{2p+p'}$, giving the following orders of magnitude for all those terms:

  $$N^{2p}n^{2p}, \text{ and } \forall p' \in \{1, \dots, p\},\ N^{2p+p'}n^{2(p-2p')},$$

  so that if N behaves like n^2, as expected, the maximum order is achieved for the first type of terms. A precise numbering of those terms of type $\{2, \dots, 2\}$ can be achieved, in a way quite similar to [Par98].

  For those terms that weigh the most among the ones of type $\{2, \dots, 2\}$, we have to choose $2p$ distinct integers in $\{1, \dots, N\}$, hence $N!/(N-2p)!$ possibilities, we also have to choose p pairs of integers in $\{1, \dots, 2p\}$, yielding $(2p)!/2^p$ possibilities. Each of those pairs is provided with a couple of the afore-mentioned integers, and each pair is either of type $\{(i, j), (i, j)\}$, $\{(i, j), (j,i)\}$, $\{(j,i), (i, j)\}$, or $\{(j,i), (j,i)\}$, each with the same approximate value $n^2/(\pi r^2)$.

This gives a total amount of $(N!(2p)!)/((N-2p)!2^p)$, that should be divided by $2^p p!$ to avoid repetitions of the same terms. Thus the sum of those terms becomes

$$\sum_{t_{(\mathbf{i},\mathbf{j})}=\{2,\ldots,2\}} E\left[\prod_{s=1}^{k} G^{(n)}_{i_s,j_s}\right] = \frac{N!(2p)!}{(N-2p)!2^p 2^p p!}\left(\frac{4n^2}{\pi r^2}\right)^p + \text{smaller terms.}$$

Let us assume that we have proven that the other terms are really negligible, then we can state

$$\begin{aligned}
&E\left[\left(a^P_{AA}(D^n)-1\right)^{2p}\right]\\
&= \sum_{N\geq 2}\frac{1}{(N(N-1))^{2p}}E\left[\left(\sum_{i,j=1}^{N} G^{(n)}_{i,j}\right)^{2p}\right]\frac{(\lambda_A n^2)^N}{N!}\exp(\lambda_A n^2),\\
&\simeq \sum_{N\geq 2}\frac{1}{(N(N-1))^{2p}}\frac{N!(2p)!}{(N-2p)!2^p 2^p p!}\left(\frac{4n^2}{\pi r^2}\right)^p\frac{(\lambda_A n^2)^N}{N!}\exp(\lambda_A n^2),\\
&\simeq \frac{(2p)!}{2^p p!}\left(\frac{2}{\pi r^2}\right)^p n^{2p} E\left[\frac{1}{N_n^{2p}}\right],
\end{aligned}$$

where we recall that N_n is Poisson distributed with mean $\lambda_A n^2$, this yields the desired asymptotics

$$\lim_{n\rightarrow+\infty} N_n^p E\left[\left(a^P_{AA}(D^n)-1\right)^{2p}\right] = \frac{(2p)!}{2^p p!}\left(\frac{2\lambda_A}{\pi r^2}\right)^p.$$

Let us show that the other terms are indeed negligible: the computation for a general term is rather tedious, so we shall only give a sketch of the proof for classes $(\mathbf{i},\mathbf{j})_w$ of cardinality 3. Firstly we will neglect the boundary effects by replacing $p_{r/n}(Y)$ by the constant $q_{r/n} = \pi r^2/n^2$ almost surely. As

$$\begin{aligned}
&(\mathbf{1}_{|Y_i-Y_j|\leq r/n} - q_{r/n})(\mathbf{1}_{|Y_k-Y_l|\leq r/n} - q_{r/n})(\mathbf{1}_{|Y_s-Y_t|\leq r/n} - q_{r/n})\\
&= \mathbf{1}_{|Y_i-Y_j|\leq r/n}\mathbf{1}_{|Y_k-Y_l|\leq r/n}\mathbf{1}_{|Y_s-Y_t|\leq r/n}\\
&-q_{r/n}\Big(\mathbf{1}_{|Y_i-Y_j|\leq r/n}\mathbf{1}_{|Y_k-Y_l|\leq r/n} + \mathbf{1}_{|Y_k-Y_l|\leq r/n}\mathbf{1}_{|Y_s-Y_t|\leq r/n} +\\
&\mathbf{1}_{|Y_i-Y_j|\leq r/n}\mathbf{1}_{|Y_s-Y_t|\leq r/n}\Big)\\
&+q^2_{r/n}\Big(\mathbf{1}_{|Y_i-Y_j|\leq r/n} + \mathbf{1}_{|Y_k-Y_l|\leq r/n} + \mathbf{1}_{|Y_s-Y_t|\leq r/n}\Big) - q^3_{r/n},
\end{aligned}$$

we may sum up the number of such terms and their values in the following table:

Type of term	Number of terms	Value
$((1,2))((2,3))((3,4))$	$\sim N^4$	0
$((1,2))((2,3))((3,1))$	$\sim N^3$	0
$((1,2))((2,3))((3,2))$	$\sim N^3$	0
$((1,2))((1,2))((2,3))$	$\sim N^3$	0
$((1,2))((1,2))((1,2))$	$\sim N^2$	$\sim q^{-2}_{r/n}$

Table 2 Enumeration of the different terms of cardinality 3 and their respective values.

Hence those terms contribute at most for N^2n^4. Terms of type for instance $\{3,3,2,\ldots,2\}$ contribute to an amount of at most $(N^2n^4)^2N^{2q}n^{2q}$, where $2p = 6+2q$, yielding an order of magnitude $N^{2p-2}n^{2p+2}$ *versus* $N^{2p}n^{2p}$ for the terms $\{2,\ldots,2\}$, recalling that the right order for N is n^2 shows the negligibility of those terms.

- If $k = 2p+1$, on may generalize the asymptotics above to conclude that the moment rescaled by the factor $N_n^{p+1/2}$ tends to 0 in a direct, though technical, way.

This conclude the proofs as the moments converge to the moments of the Normal law.