# Is the radial distance really a distance? An analysis of its properties and interest for the matching of polygon features

Yann Méneroux, Ibrahim Maidaneh Abdi, Arnaud Le Guilcher, Ana-Maria Olteanu-Raimond

# Is the *radial distance* really a distance? An analysis of its properties and interest for the matching of polygon features.

Yann Méneroux[a], Ibrahim Maidaneh Abdi[a,b], Arnaud Le Guilcher[a] and Ana-Maria Olteanu-Raimond[a]

[a] LASTIG, Univ Gustave Eiffel, ENSG, IGN, F-94160 Saint-Mandé, France; [b] ITU-I, the University of Djibouti, Djibouti

**ARTICLE HISTORY**
Compiled September 3, 2022

**ABSTRACT**
In this paper we examine the properties of the *radial distance* which has been used as a tool to compare the shape of simple surfacic objects. We give a rigorous definition of the *radial distance* and derive its theoretical properties, and in particular under which conditions it satifies the distance properties. We show how the computation of the *radial distance* can be implemented in practice, and made faster by the use of an analytical formula and a Fast Fourier Transform. Finally, we conduct experiments to measure how the *radial distance* is impacted by perturbation and generalization and we give abacuses and thresholds to deduce when buildings are likely to be homologous or non-homologous given their *radial distance*.

**KEYWORDS**
Radial distance; shape similarity; polygon data matching; robustness; error propagation

## 1. Introduction

Geographic information (GI) production practices have changed considerably in recent decades. Open data policies as well as technological advances have had a significant impact on the way GI is collected, stored, analyzed and disseminated. Whether at European, national or regional level, different public actors produce and share GI. In addition, the development of low-cost location-based devices and the rise of Web 2.0 paradigm initiatives have led to a considerable mass of GI and changed citizens' perceptions of its acquisition and use. This alternative GI produced by citizens is communally known as volunteered geographic information (Goodchild 2007). Among all VGI initiatives (See *et al.* 2016), one of the them is the well known OpenStreetMap (OSM) which started from a strong desire to open up GI and offer a data model with high flexibility. In this context, many sources of GI describing the territory exist and are characterized by spatial, temporal, thematic and semantic heterogeneity.

There is much literature studying the use of traditional and alternative GI sources for various purposes such as validation (See *et al.* 2016, Olteanu-Raimond *et al.*

CONTACT Yann. Méneroux. Email: yann.meneroux@ign.fr

2020), update (Ivanovic *et al.* 2019b, Liu *et al.* 2021, van Winden *et al.* 2016) and enrichment of authoritative GI (Zielstra and Hochmair 2011, Al-Bakri and Fairbairn 2012), production of a new spatial repository for rescue applications (Van Damme *et al.* 2019), land use and land cover monitoring (Fonte *et al.* 2017, Schultz *et al.* 2017). Studying the quality of alternative GI by comparing it with a reference data source is a standalone research topic or just the first step of the combination of traditional and alternative GI (Fonte *et al.* 2015, Senaratne *et al.* 2017, Ivanovic *et al.* 2019a, Yan *et al.* 2020).

One of the main challenges in using both traditional and alternative GI is to identify the redundant GI (*e.g.* the same entity from the real world — a building — is represented in two data sources) and the complementary GI (*i.e.* a source of GI has a higher exhaustivity than another for the same type of feature). To cope with this issue, data matching is a powerful and very used approach (Walter and Fritsch 1999, Mustière and Devogele 2008, Fan *et al.* 2014, Olteanu-Raimond *et al.* 2015, Liu *et al.* 2015, Costes and Perret 2019, Ivanovic *et al.* 2019b). Data matching consists in defining homologous features from different sources of GI representing the same entity from the real world (*e.g.* a feature $A_1$ from a source $S_1$ is homologous with a feature $A_2$ from a source $S_2$). Many data matching algorithms exist in the research literature and they rely on the definition of similarity measures based on different feature characteristics such as: geometric positions, shapes, thematic and semantic information as well as spatial relationships, or a combination of all those measures. Generally, the data matching process is composed of three main steps. First, for each feature from a dataset, the algorithm is looking for candidates to match. Second, for each couple (selected feature, candidate) criteria based on similarity measures or topology are computed. Finally, a decision is made based on the combination of the used criteria. Note that, similarity measures and the combinations approaches depend on the type of the date (*e.g.* point, lines, polygons) or scale of the datasets to match (e.g. same or different scales). A detailed state of the art of the data matching algorithms and similarity measures is described in Olteanu-Raimond *et al.* (2015) and Xavier *et al.* (2016).

When dealing with heterogeneous GI coming from multisources such as different data models, mapping specifications, spatial and temporal scales, as well as non exhaustive or incomplete semantic and thematic information in data sources to be matched, such as for building data, deriving similarity measures that are robust to complex polygons and geometric errors becomes a challenge. Shape measures proved to be efficient for matching polygons datasets since they better distinguish the features between them (Meng and Lu 2014, Fan *et al.* 2014, Xavier *et al.* 2016, Maidaneh Abdi *et al.* 2020). Thus, in the context of heterogeneous GI data matching, there is a need to describe the similarity of polygons in a very fine way.

One of the shape measures is the *radial distance* which was initially proposed by Cohen and Guibas (1997) and successfully used for matching polygon features (Vauglin and Bel Hadj Ali 1998, Maidaneh Abdi *et al.* 2020) or for building footprint squaring operations (Lokhat and Touya 2016). It describes the shape of a polygon feature by measuring the distances from the centroid of the polygon to the points composing its boundary. Despite the success of its use and its properties to measure the similarity between two polygons, as mentioned by Hangouët (2006), the mathematical definition of the *radial distance* still needs to be formalized. Moreover, in the research proposed by Basaraner and Cetinkaya (2017), among 20 studied indices, 2 of them are based

on radial signature (*i.e.* shape membership index and roughness index). Thus, it is relevant to study radial distance also based on radial signature. To the authors' knowledge, there is still no mathematical proof that the radial distance is a distance.

The aim of this paper is to propose mathematical definitions for both the radial signature and the *radial distance*, to demonstrate under which conditions the *radial distance* is a distance and to study its robustness from different perspectives in the context of heterogeneous polygon data matching.

The paper is organized as follows. Section 2 describes the state of the art regarding shape similarity measures and the limitations of the *radial distance*. Section 3 introduces the formal definitions and properties associated to the *radial distance*. Section 4 details some proposals to fasten the computation of the *radial distance*. The robustness of the *radial distance* is studied in Section 5 followed, in Section 6, by an experimentation on real data stemming from two buildings datasets: authoritative open data produced by IGN France and OpenStreetMap. Finally, conclusions and future work are outlined in Section 7.

## 2. State of the art

Depending on whether the boundary or the core of the polygon is considered, the shape measures are categorized by Zhang and Lu (2004) into two groups: boundary based and core based. Basaraner and Cetinkaya (2017) also described an exhaustive study analyzing the performance of 20 shape measures with respect to the complexity of building footprints. Some are commonly used such as circularity, convexity, fractality, and some are less frequent measures such as squareness, shape memberships or measures implemented in the Shape Metrics Toolbox (Angel *et al.* 2010). Hereafter, we briefly describe the shape similarity measures used in data matching.

For complex polygons with holes or multi-polygons, core based methods are used. One of the most used approaches are those stemming from image matching or feature recognition fields (Adoram and Lew 1999, Kim and Kim 2000, Premaratne and Premaratne 2014). In the field of GIS, Fu *et al.* (2010) recently proposed a new shape similarity measure mixing Fourier transform to derive the boundary and local moments to extract the core of the polygons. The shape measure is invariant to translation, scale and rotation.

For simple polygons, boundary based approaches are used. The turning function describes a polygon feature by measuring the angles formed by the segments composing its boundary and a horizontal half-line oriented along the abscissa axis (Arkin *et al.* 1991, Fu *et al.* 2010). A turning signature is then computed and normalized by the perimeter of the polygon feature. The shape measure based on angles is invariant to translation, scale and rotation. When a rotation $\alpha$ is applied to the polygon, its corresponding turning function is translated along the $y$-axis. To compare two turning functions, finding an optimal shift is required.

Finally, another category of boundary based approaches consider the centroid of polygon, which ensure the translation invariance. For example, Chang *et al.* (1991)

defines a distance function by computing the distances between the polygon centroid and vertices. The computed distances are ordered and normalized by the minimal distance to achieve both rotation and scale invariance. The *radial distance* is part of this category. In order to be considered as a similarity measure for data matching, in addition to be a metrics, the *radial distance* must be invariant to translation, rotation and scaling, easy to compute and interpret (Arkin *et al.* 1991). We can add to these characteristics the robustness which is absolutely necessary when dealing with heterogeneous GI.

In this context, the contributions of our paper are:

- a mathematical definition of the *radial distance*,
- a theoretical demonstration of its properties, and the theoretical and empirical conditions under which they are valid,
- a theoretical analysis of the impact of a one-off disturbance,
- a formalization of the algorithm allowing to compute the *radial distance* and two optimizations to reduce the computation time for complex polygons,
- a set of experiments on real data, characterizing the impact on the *radial distance* of two types of the most common noise in geographic databases, as well as recommendations and criteria for deciding whether two buildings are potentially homologous given their respective shapes. The variety of the proposed criteria allows to adapt to different contexts, with different prior knowledge.

## 3. Definition and properties of the *radial distance*

Defining distances between subsets of $\mathbb{R}^2$ such as polygons or curves is not always natural, and one can conceive different distances when emphasizing different aspects such as position, orientation or shape. The *radial distance* focuses on shape.

### Definition 3.1. Closed curve

A non-self-intersecting closed curve $\widetilde{\gamma}$ of length $L$ is the image in $\mathbb{R}^2$ of any injective function $\gamma$:

$$\begin{array}{rccc} \gamma: & [0,L] & \rightarrow & \mathbb{R}^2 \\ & s & \mapsto & \gamma(s), \end{array}$$

with $\gamma(0) = \gamma(L)$. We note $I(\gamma) \subset \mathbb{R}^2$ the compact delimited by $\gamma$, and $C(\gamma)$ the centroid of $I(\gamma)$. We suppose that $\gamma$ turns anticlockwise (if we define an orientation of $\gamma$ by ascending curvilinear abscissa, $I(\gamma)$ is at the left of $\gamma$).

In general, when it is clear from the context, we will identify the function $\gamma$ and its image $\widetilde{\gamma} = \mathrm{Im}(\gamma)$. We note $\Gamma$ the set of all curves defined by Definition 3.1.

### Definition 3.2. Radial signature

The radial signature of a curve $\gamma \in \Gamma$ is the function defined by:

$$\begin{array}{rccc} r_\gamma : & [0,1] & \to & \mathbb{R} \\ & t & \mapsto & \dfrac{1}{R}\|\gamma(tL) - C(\gamma)\| \end{array}$$

with $R = \|\gamma(\cdot L) - C(\gamma)\|_{L^2}$ so that $\forall\, \gamma \in \Gamma \ : \ \|r_\gamma\|_{L^2} = 1$, where $\|\cdot\|_{L^2}$ is the canonical norm of the $L^2([0,1])$ space.

We note that $r_\gamma$ satisfies $r_\gamma(0) = r_\gamma(1)$. In the following, we will also use the notation $r_\gamma$ for the function defined on $\mathbb{R}$ which is 1-periodic and coincides with $r_\gamma$ on $[0,1]$.

**Definition 3.3. Radial distance**

The *radial distance* between two closed, non self-intersecting, anticlockwise curves $\gamma$ and $\gamma'$ is defined by

$$d_r^2(\gamma, \gamma') = \min_{\tau \in [0,1]} \int_0^1 \left(r_\gamma(t) - r_{\gamma'}(t+\tau)\right)^2 dt$$

Figure 1 illustrates an example of radial signatures and *radial distance* computation on two simple building polygons. Figure 1 a. depicts two similar polygons, defined by closed curves $\gamma_A$ (red) and $\gamma_B$ (blue). First, radial signatures $r_A$ and $r_B$ are evaluated (see figure 1 b.) as functions of the radial distances of polygon boundary points to the centroid (in m), versus curvilinear abscissa (renormalized between 0 and 1). These signatures are similar and need to be registered against each other in order to compute the *radial distance*. Figure 1 c. depicts the $L^2$-norm of the function $t \mapsto r_A(t) - r_B(t+\tau)$, for all possible abscissa shifts $\tau$. This step is essential since it makes the radial distance invariant to the starting points of polygons. The shift resulting in a minimal $L^2$-norm is located at the point $\tau = 17\%$. This amounts to apply a 0.17 circular shift to the left on the blue radial signature, to make it correspond to the red signature, as depicted on figure 1 d.

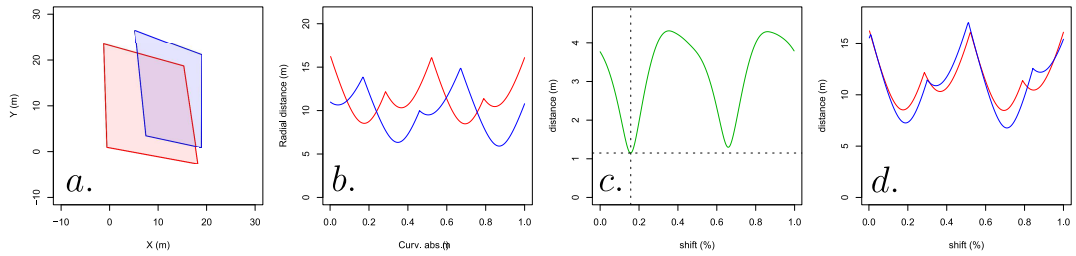


**Figure 1.** Example of radial signatures and *radial distance* for two polygons: a. two polygons $\gamma_A$ (red) and $\gamma_B$ (blue); b. their respective (non-normalized) radial signatures; c. the distance between radial signatures with respect to shift; d. the radial signature with optimal shift (17 %).

Despite being often called *radial distance*, we will show hereafter that it is technically not a distance, since it is not guaranteed that two different polygons are separated by a strictly positive distance (Proposition 3.6). However, we demonstrate that it is

a pseudometrics (Proposition 3.5), meaning that all other properties of a distance are guaranteed, and in particular the triangular inequality. Furthermore, we will demonstrate that it is at least a distance on a subset of polygons which are naturally encountered in building shapes (Proposition 3.7). First, let us start by recalling the definition of a distance.

**Definition 3.4. Distance**

A function $d : \Gamma \times \Gamma \to \mathbb{R}_+$ is a distance, if and only if, given any three elements $\alpha$, $\beta$, $\gamma \in \Gamma$, $d$ satisfies the three following properties :

- (i) $d_r(\alpha, \beta) = 0 \Leftrightarrow \alpha = \beta$
- (ii) $d_r(\alpha, \beta) = d_r(\beta, \alpha)$
- (iii) $d_r(\alpha, \gamma) \leqslant d_r(\alpha, \beta) + d_r(\beta, \gamma)$

The *radial distance* being a shape measure, in the subsequent theoretical developments, we consider two polygons as being equal, if it is possible to find a direct affine conformal transformation of $\mathbb{R}^2$ between them. Hence, we implicitly work in the quotient set $\Gamma/\sim$, where $\sim$ is the equivalence relation denoting the existence of such a transformation between the two polygons.

**Proposition 3.5. Symmetry and triangular inequality**

*The function $d_r : \Gamma \times \Gamma \to \mathbb{R}_+$ is a pseudometrics, i.e. given any three closed curves $\alpha$, $\beta$, $\gamma \in \Gamma$, $d_r$ satisfies the three following properties :*

- *(i')* $d_r(\gamma, \gamma) = 0$
- *(ii)* $d_r(\alpha, \beta) = d_r(\beta, \alpha)$
- *(iii)* $d_r(\alpha, \gamma) \leqslant d_r(\alpha, \beta) + d_r(\beta, \gamma)$

Note that $(i')$ is only the direct and weak version of the separation property $(i)$ required for a full metrics.

**Demonstration**

*(i')* It is clear that, being the minimum of a positive function, $d_r(\gamma, \gamma)$ is also positive. Besides, choosing $\tau = 0$ results in $\int_0^1 \big(r_\gamma(t) - r_\gamma(t+\tau)\big)^2 dt = \int_0^1 \big(r_\gamma(t) - r_\gamma(t)\big)^2 dt = 0$. As a consequence : $d_r(\gamma, \gamma) \geqslant 0$ and $d_r(\gamma, \gamma) \leqslant 0 \Rightarrow d_r(\gamma, \gamma) = 0$

*(ii)* Successively by integral variable substitution $t \mapsto t - \tau$, symmetry of the square function and then 1-periodicity of the radial signatures:

$$\int_0^1 \big(r_\beta(t) - r_\gamma(t+\tau)\big)^2 dt = \int_\tau^{1+\tau} \big(r_\beta(t-\tau) - r_\gamma(t)\big)^2 dt = \int_\tau^{1+\tau} \big(r_\gamma(t) - r_\beta(t-\tau)\big)^2 dt$$
$$= \int_0^1 \big(r_\gamma(t) - r_\beta(t-\tau)\big)^2 dt$$

As a consequence, we have the following equality:

$$\forall\ \tau \in [0,1] \quad \int_0^1 \big(r_\beta(t) - r_\gamma(t+\tau)\big)^2 dt = \int_0^1 \big(r_\gamma(t) - r_\beta(t-\tau)\big)^2 dt$$

$$\begin{aligned}
\Rightarrow d_r(\beta,\gamma) = \min_{\tau\in[0,1]} \int_0^1 \big(r_\beta(t) - r_\gamma(t+\tau)\big)^2 dt &= \min_{\tau\in[0,1]} \int_0^1 \big(r_\gamma(t) - r_\beta(t-\tau)\big)^2 dt \\
&= \min_{\tau\in[0,1]} \int_0^1 \big(r_\gamma(t) - r_\beta(t+1-\tau)\big)^2 dt \\
(1-\tau \mapsto \tau') \qquad &= \min_{\tau'\in[0,1]} \int_0^1 \big(r_\gamma(t) - r_\beta(t+\tau')\big)^2 dt \\
&= d_r(\gamma,\beta)
\end{aligned}$$

*(iii)* We note $r_\alpha$, $r_\beta$ and $r_\gamma$ the radial signatures associated to closed curves $\alpha$, $\beta$ and $\gamma$ respectively, and $d(r,s) = \|r-s\|_{L^2}$.

Let $G$ be a group and $\varphi$ a group action of $G$ on $X = L^2([0,1])$ defined by $\varphi : G \times X \to X$ such that $\varphi(\mathbf{e}, g) = g$ for any $g \in X$ if $\mathbf{e}$ is the identity element of $G$ and $\varphi(\mathbf{e_1}, \varphi(\mathbf{e_2}, f)) = \varphi(\mathbf{e_1} * \mathbf{e_2}, f)$ where $*$ is the binary operator on $G$. We show that, if $d(f,g) = d(\varphi(\mathbf{x}, f), \varphi(\mathbf{x}, g))$ for any element $\mathbf{x} \in G$, then the application:

$$\begin{aligned}
d_r : X \times X &\to \mathbb{R}_+ \\
f, g &\to \min_{\mathbf{x},\mathbf{y}\in G}\ d\big(\varphi(\mathbf{x}, f), \varphi(\mathbf{y}, g)\big)
\end{aligned}$$

is a distance on $X$.

First, we note that since $G$ is a group, for every element $\mathbf{x} \in G$ there exists $\mathbf{x}^{-1}$ $(\in G)$ such that $\mathbf{x}^{-1} * \mathbf{x} = \mathbf{x} * \mathbf{x}^{-1} = \mathbf{e}$, and the definition of $d_r$ may be rewritten as:

$$\begin{aligned}
d_r(f,g) &= \min_{\mathbf{x},\mathbf{y}\in G}\ d\big(\varphi(\mathbf{x}, f), \varphi(\mathbf{y}, g)\big) = \min_{\mathbf{x},\mathbf{y}\in G}\ d\big(\varphi(\mathbf{x}^{-1}, \varphi(\mathbf{x}, f)), \varphi(\mathbf{x}^{-1}, \varphi(\mathbf{y}, g))\big) \\
&= \min_{\mathbf{x},\mathbf{y}\in G}\ d\big(\varphi(\mathbf{x}^{-1} * \mathbf{x}, f), \varphi(\mathbf{x}^{-1} * \mathbf{y}, g)\big) = \min_{\mathbf{z}\in G}\ d\big(\varphi(\mathbf{e}, f), \varphi(\mathbf{z}, g)\big) \\
&= \min_{\mathbf{z}\in G}\ d\big(f, \varphi(\mathbf{z}, g)\big)
\end{aligned}$$

with $\mathbf{z} = \mathbf{x}^{-1} * \mathbf{y}$. This alternative formulation may be identified with the definition of $d_r$ in Definition 3.3, with $\varphi$ the parameterization shift operation, $\mathbf{z} = +\tau$ the shift,

and $g = r_{\gamma'}$, one of the two radial signatures to compare.

The separation and symmetry of $d_r$ clearly result from the fact that $d$ itself is a distance. Let us consider the orbits $\mathcal{O}_\alpha$, $\mathcal{O}_\beta$, and $\mathcal{O}_\gamma$ of $\alpha$, $\beta$, and $\gamma$ respectively, where $\mathcal{O}_x = \{y \in X \mid \exists\, g \in G : y = \varphi(g, x)\}$.

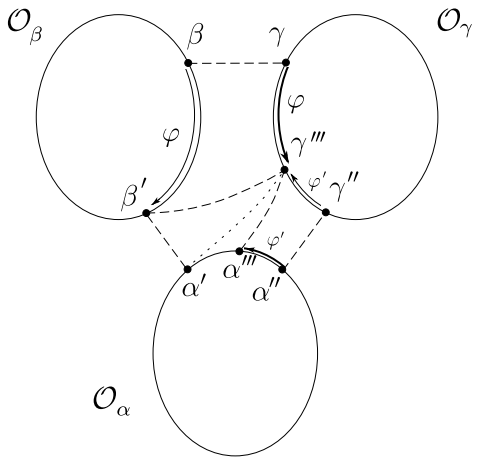


**Figure 2.** Triangular inequality of the quotient distance $d_r$ between the orbits of three radial signatures.

As depicted on Figure 2, we note $\beta$ and $\gamma$ the optimal elements pertaining to $\mathcal{O}_\beta$ and $\mathcal{O}_\gamma$ respectively, and minimizing $d(\beta, \gamma)$. Similarly, we note $\alpha'$ and $\beta'$ the elements of $\mathcal{O}_\alpha$ and $\mathcal{O}_\beta$ minimizing $d(\alpha, \beta)$. Eventually, $\alpha''$ and $\gamma''$ are the elements of $\mathcal{O}_\alpha$ and $\mathcal{O}_\gamma$ minimizing $d(\alpha, \gamma)$.

Since, $\beta, \beta' \in \mathcal{O}_\beta$, there exists a group action element $\varphi$ transforming $\beta$ into $\beta'$ and such that $d(\beta, \gamma) = d(\beta', \gamma''')$ where $\gamma'''$ is an arbitrary element of the orbit $\mathcal{O}_\gamma$. Similarly, there is a group action element $\varphi'$ transforming $\gamma''$ into $\gamma'''$ and such that $d(\alpha'', \gamma'') = d(\alpha''', \gamma''')$ where $\alpha'''$ is an arbitrary element of the orbit $\mathcal{O}_\alpha$. Let us consider the triangle $\alpha'\beta'\gamma'''$. Since $d$ is a distance on $X$ the triangular inequality states that : $d(\alpha', \gamma''') \leqslant d(\alpha', \beta') + d(\beta', \gamma''')$. Since $(\alpha'', \gamma'')$, and then by isometric transformation $\varphi'$ $(\alpha''', \gamma''')$, are supposed to minimize the distance $d$ between the elements of $\mathcal{O}_\alpha$ and $\mathcal{O}_\gamma$ we have $d(\alpha''', \gamma''') \leqslant d(\alpha', \gamma''')$. It follows that:

$$d(\alpha''', \gamma''') \leqslant d(\alpha', \beta') + d(\beta', \gamma''') \quad \Leftrightarrow \quad d(\alpha'', \gamma'') \leqslant d(\alpha', \beta') + d(\beta, \gamma)$$

$$\Leftrightarrow \min_{\mathcal{O}_\alpha, \mathcal{O}_\gamma} d(\alpha, \gamma) \leqslant \min_{\mathcal{O}_\alpha, \mathcal{O}_\beta} d(\alpha, \beta) + \min_{\mathcal{O}_\beta, \mathcal{O}_\gamma} d(\beta, \gamma)$$

$$\Leftrightarrow \quad d_r(\alpha, \gamma) \leqslant d(\alpha, \beta) + d(\beta, \gamma)$$

It is immediate to show that the translation $\tau$ of the origin point of a closed curve is an isometric action group. As a consequence, $d_r$ satisfies the triangular inequality on the set $\Gamma$ of closed curves. ■

Then $d_r$ is a pseudometrics on $\Gamma$. It would be a full metrics if the separation was also guaranteed which is not the case (see following proposition). However, in the special case of polygon comparison for geographic data matching, we may argue that triangular inequality is the most important property that must be satisfied. As illustrated on Figure 3, let us consider two databases of polygons, containing $\gamma$ and $\gamma'$ respectively, that may (or may not) represent the same feature in the *real world*. Further, let us imagine that in the second database $\gamma'$ is perturbated by a noise process or has been generalized, resulting in a slightly modified version $\gamma''$ of $\gamma'$. Suppose also that there is some upper bound guarantee on this perturbation process, for example that it exists a fixed (known) quantity $\varepsilon \in \mathbb{R}_+$ such that $d_r(\gamma', \gamma'') \leqslant \varepsilon$. According to the triangular inequality property of $d_r$, $d_r(\gamma, \gamma') \leqslant d_r(\gamma, \gamma'') + d_r(\gamma'', \gamma')$ and then, the reverse triangular inequality guarantees that :

$$\Delta \equiv \left| d_r(\gamma, \gamma') - d_r(\gamma, \gamma'') \right| \leqslant d_r(\gamma'', \gamma') \leqslant \varepsilon$$

In other words, the impact of the noise or of the generalization process on the features comparison with $d_r$ is no larger than the amplitude of the perturbation itself.

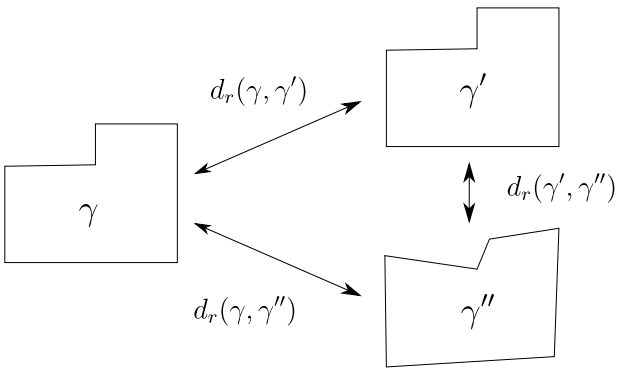


**Figure 3.** Two polygons being compared with the pseudometrics $d_r$ : on one hand $\gamma$ contained in one database and on the other hand $\gamma'$ or its noised or generalized version $\gamma''$ in another database. Any upper bound on the perturbation effect will automatically bound the error $\Delta$ on $d_r$ comparison between the two polygons.

Similarly, as illustrated on Figure 4, for a more general data collection pipeline, we may consider three distinct components: the topographic database specifications $\varepsilon$ (which leads to different *nominal terrains*, implying that a same object in the *real world* may be represented by two different shapes in the databases, even with perfect measurement and no further data post-processing), the measurement noise $\varepsilon'$ which is inherent to any survey operation and eventually any perturbation $\varepsilon''$ added in post-processing steps, such as cartographic generalization. Since, $d_r$ enforces triangular inequality, it is known that the distance between the *real world* feature and its polygonal representation in a topographic database, is at most equal to the sum of individual error terms: $\varepsilon + \varepsilon' + \varepsilon''$.

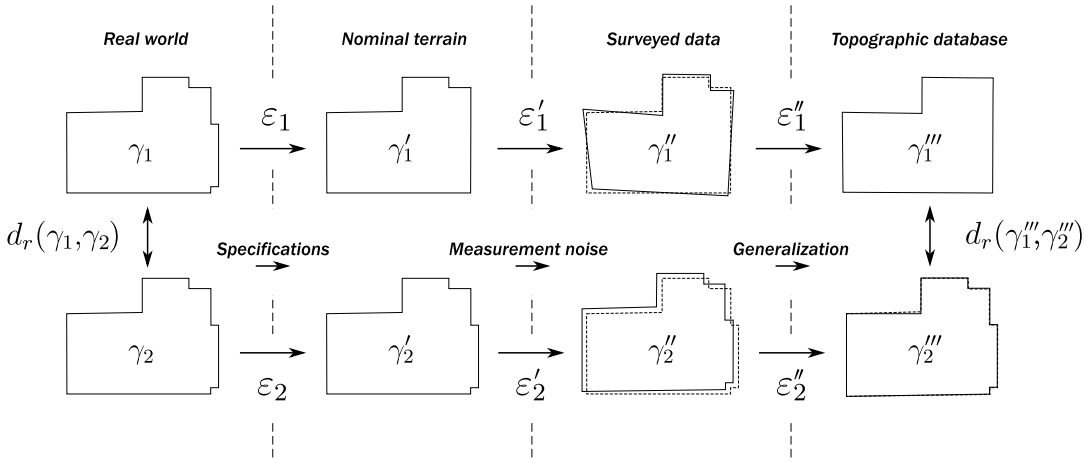


**Figure 4.** Two polygons being compared with the pseudometrics $d_r$ within two different data collection pipeline. Each pipeline ($i = 1, 2$) is perturbated by potentially different specifications ($\varepsilon_i$), different measurement noise ($\varepsilon'_i$) and different generalization process ($\varepsilon''_i$).

When comparing two different topographic databases (indexed by $i = 1, 2$), the perturbation process may be different for each database, hence we note : $\varepsilon_i$, $\varepsilon'_i$ and $\varepsilon''_i$. However, triangular inequality still ensures that the error in distance evaluation (before and after data collection process) is bounded by the sum of individual components:

$$\Delta \equiv \left| d_r(\gamma_1, \gamma_2) - d_r(\gamma'''_1, \gamma'''_2) \right| \leqslant \varepsilon_1 + \varepsilon_2 + \varepsilon'_1 + \varepsilon'_2 + \varepsilon''_1 + \varepsilon''_2$$

Consequently, if there is a control on each of the individual perturbation processes $\varepsilon$, the distance evaluation error $\Delta$ may be bounded. In the particular case where the two polygons represent the same entity in the *real world*, $d_r(\gamma_1, \gamma_2) = 0$, and its evaluation $d_r(\gamma'''_1, \gamma'''_2)$ based on the surveyed data is bounded.

**Proposition 3.6.** $d_r$ **is not separated**. *It is possible to find two distinct polygons* $\gamma, \gamma' \in \Gamma$*, with* $d_r(\gamma, \gamma') = 0$.

In the general case, $d_r$ does not satisfy separation, *i.e.* $d_r(\gamma, \gamma') = 0$, does not necessarily implies that $\gamma = \gamma'$. A counter-example is provided in Figure 5.

Given an arbitrary centroid $O \in \mathbb{R}^2$, $A(1, 0)$, $B(1, 1)$, $D(0.3, 0.3)$, $E(0.3, 1)$, and $A'(0, 1)$ are set. Note that $O$, $B$ and $D$ are aligned. Then, $C$ is sampled at an arbitrary position within the triangle $BDE$ and $C'$ defined as its symmetric with respect to the $\mathbf{y} = \mathbf{x}$ axis. Then, a (non-closed) curve $\alpha$ is defined by the sequence of vertices $ABCDEA'$ (note that $C'$ is not included in this curve). The closed curve $\gamma$ is defined by symmetry as illustrated on Figure 5. It is not difficult to show that $\gamma$ is indeed a non-intersecting curve enclosing an area whose centroid $C(\gamma)$ is $O$ by construction.

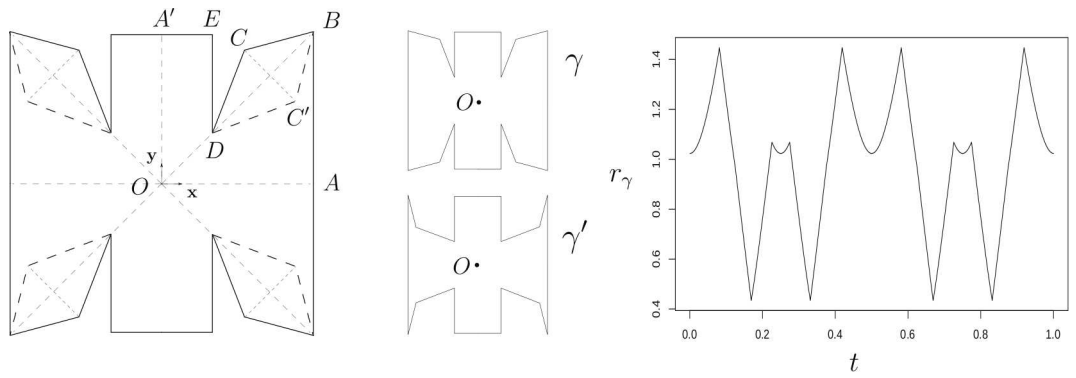


**Figure 5.** A counter-example for the separation illustrating that $d_r$ is not a distance: a curve $\gamma$, and its alternative form $\gamma'$ such that $\gamma \neq \gamma'$, and yet $d_r(\gamma, \gamma') = 0$

Similarly, we define a second curve $\gamma'$ following the same steps as above, except that the vertex $C'$ is substituted to $C$. We also have $C(\gamma') = C(\gamma) = O$.

Clearly $\gamma \neq \gamma'$. However, arbitrarily parameterizing $\gamma$ and $\gamma'$ with $s = 0$ at vertex $A$, both curves are perfectly symmetric with respect to their common centroid $O$, regarding radial coordinate: $\forall\ s \in [0, L]\ ||C(\gamma) - \gamma(s)|| = ||C(\gamma') - \gamma'(s)||$ at every point $s$, which in turns implies that $r_\gamma = r_{\gamma'}$ and then $d_r(\gamma, \gamma') = 0$. ■

**Proposition 3.7. Separation for star-shaped polygons** *We consider $\Gamma^*$ the set of polygons $\gamma$ for which $I(\gamma)$ is star-shaped around its centroid $C(\gamma)$ (satisfying, for all $s \in [O, L], t \in [0, 1[$, $(1-t)C(\gamma) + t\gamma(s) \in I(\gamma)$). Then the radial distance satisfies the following property:*

$$\forall\ \gamma, \gamma' \in \Gamma^*, \quad d_r(\gamma, \gamma') = 0 \Rightarrow \gamma = \gamma' \tag{1}$$

*where the equality is to be understood up to any direct similarity map of $\mathbb{R}^2$.*

To describe the principle of the proof, we can associate with each point its linear abscissa $s$ (which is such that $s = tL$), and its polar coordinates $(\rho, \varphi)$ when the centroid of the polygon is taken as the origin. If one considers only polygons that are star-shaped around their center of mass, the angle $\varphi$ is a non-decreasing function of $t$, and counter-examples like the one exhibited in Proposition 3.6 do not exist. Now if the length of the curve $L$ is known, the function $s \mapsto \rho(s)$ defines a unique polygon (star-shaped around its center of mass) up to an isometry. Indeed, the polygon can be reconstructed iteratively, edge after edge, and for each edge the position of the next vertex is uniquely determined by the function $\rho(s)$. Because the radial signature is obtained after two re-normalizations, and we do not know *a priori* whether and how these two re-normalizations are related, the demonstration is less straightforward.

The key of the proof is the observation that $\varphi(t = 1)$ must be equal to $\varphi(t = 0) + 2\pi$ (because we consider non-self-intersecting polygons that are oriented anti-clockwise). This equality actually sets the value of the ratio between the two re-normalization factors $L$ and $R = \frac{\rho(\frac{\cdot}{L})}{r}$, from which we can infer that the radial signature corresponds to a unique star-shaped polygon (up to an homotopy). To illustrate that, we consider

an arbitrary value of $R$ and see how it determines $L$. For a given $L$, the radial signature allows the reconstruction of a unique polyline, but these polylines have all different values for $\varphi(L) - \varphi(0)$, and there is only one value $L_0 > 0$ such that $\varphi(L_0) - \varphi(0) = 2\pi$. For $L < L_0$, the polyline makes less than one turn and does not close itself ; for $L > L_0$ the polyline does more than one turn and cannot close itself without being self-intersecting. Once it is proven that the signature fixes the ratio between $R$ and $L$, it is straightforward to show that two centroid-star-shaped polygons sharing the same radial signature are equal up to an homotopy.

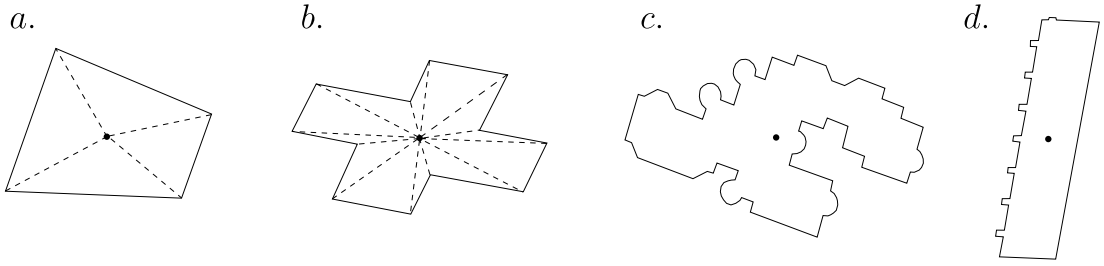


**Figure 6.** Examples of building polygons extracted from Aubervilliers (France) dataset: a. A convex polygon. b. A non-convex but yet star-shaped polygon, for which separation is still guaranteed (segments from center of mass to any point are fully contained in the polygon). c. A non-star-shaped polygon. d. A *nearly* star-shaped polygon (about 94% of its boundary, and 99.7 % of its surface is *visible* from the center of mass).

We showed that $d_r$ is strictly speaking not a distance on $\Gamma$, however, Proposition 3.7 guarantees that it can be used as such, at least for the subset $\Gamma^* \subset \Gamma$ of star-shaped polygons. This is an important result, since in practical applications, most geographic features (in particular building polygons) have nearly convex, and often star-shaped geometries. Some tests conducted on Aubervilliers city (France) building database showed that about 67 % of polygons are star-shaped. Moreover, for the remaining polygons, we evaluated the proportion of angles being not star-shaped (*i.e.* for which the semi-infinite line of sight do not contain exactly one intersection with the building perimeter), and found an average value of 68° (meaning that the polygons are on average 81 % star-shaped). Overall, 9 buildings out of 10 are more than 90 % star-shaped regarding perimeter proportion.

**Proposition 3.8. Radial distance is bounded** $Im(d_r) = d_r(\Gamma \times \Gamma) = \big[0, \sqrt{2}\big[$.

The supremum value of $d_r$ on $\Gamma \times \Gamma$ is finite and equal to $\sqrt{2}$. This is an interesting property, since it shows that *radial distance* can be converted to a similarity metrics $s : \Gamma \times \Gamma \to\ ]0, 1]$ through the transformation : $s(\gamma_1, \gamma_2) = 1 - d_r(\gamma_1, \gamma_2)/\sqrt{2}$.

Trivially, $\mathrm{Im}(d_r) \subseteq \mathbb{R}_+$. It remains to demonstrate that: $(i)$ $d_r$ is (strictly) bounded by $\sqrt{2}$ and $(ii)$ that this upper bound is sharp. We provide here a short sketch of the proof, for more theoretical considerations, refer to appendix B, where the involved constructive proof is interesting as well, insofar as it suggests that star-shape property might also nearly be a *necessary* condition for $d_r$ to be a distance.

$(i)$ With two signatures $f$ and $g$ being in the space of non-negative unit $L^2$-norm functions, we have $\|f - g\|^2_{L^2} = \|f\|^2_{L^2} + \|g\|^2_{L^2} - 2\langle f|g\rangle = 2(1 - \langle f|g\rangle)$. Since the inner product of two non-negative functions is also non-negative, it follows that:

$\|f - g\|_{L^2} \leqslant \sqrt{2}$, which provides an upper bound on the values taken by the *radial distance*. Besides, reaching this upper bound would require that $f$ and $g$ are orthonormal, which is possible only if they are both simultaneously non-zero on at most a countable subset of $[0, 1]$. Such a requirement cannot be met with continuous functions, hence there is no pair of buildings whose radial distance reaches $\sqrt{2}$.

(*ii*) To show that the upper bound is sharp, we need to find pairs of closed curves whose radial distance is arbitrarily close to it. To do so, we first notice that any unit-norm non-negative piecewise constant $\frac{1}{2}$-periodic function of $[0, 1]$ can be considered as the simple limit of a sequence of radial signatures (a constructive proof, illustrated on Figure 7, is provided in appendix B). Hence, the *double gate* function $\Pi_n$ with gates of width $1/2n^2$ and height $n$ is the limit of a sequence of radial signatures of closed curves $(\gamma_m^n)_{m\in\mathbb{N}}$, whose radial distance to the unit-circle converges to $\sqrt{2}$, which is then the supremum value of $\mathrm{Im}(\Gamma)$. ■

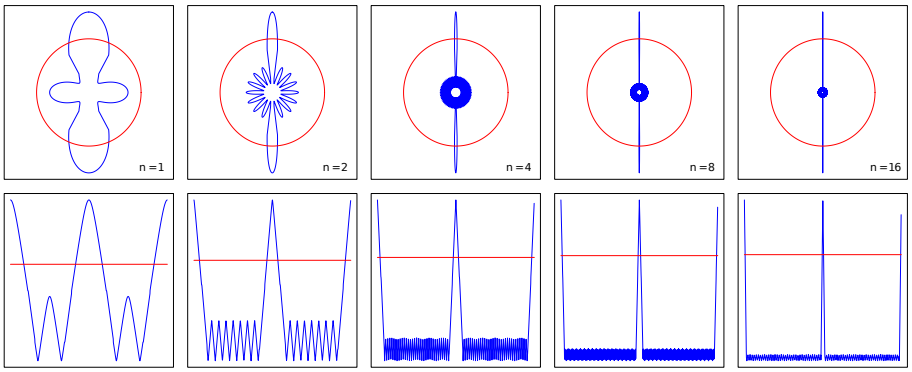


**Figure 7.** A sequence $(\gamma_n)_{n\in\mathbb{N}}$ of closed curves (top) and their respective radial signatures (bottom in blue). The *radial distance* between $\gamma_n$ and any circle (in red) tends to the supremum value $\sup(d_r(\Gamma \times \Gamma)) = \sqrt{2}$

.

However, this upper bound is a theoretical value, typical *radial distance* values encountered in practice are much lower than $\sqrt{2}$, as indicated by the experimentation results in Section 6. Further investigations are provided in Appendix B to explain this.

## 4. Methods for the computation of the *radial distance*

In this section, we provide the pseudo-code algorithm to compute the radial signatures and the *radial distance* of two polygons, given in input as a list of 2D vertices. Then we propose two possible improvements to accelerate the computation, especially when polygons are defined with a large number of vertices and/or a small discretization step in the signatures is required. The first contribution enables a fast computation of the radial signatures with infinite resolution through an interpolation based on a piecewise analytical form of the signatures (section 3.2). The second contribution offers a fast search of the optimal shift to register two radial signatures (section 3.3).

### 4.1. *Direct computation*

The first module hereafter takes as input two arrays $\mathbf{X}$ and $\mathbf{Y}$, each containing $N$ floating point numbers, where $N-1$ stands for the number of distinct vertices in the polygon. Note that for the sake of simplicity, we require that $\mathbf{X}[0] = \mathbf{X}[N]$ and $\mathbf{Y}[0] = \mathbf{Y}[N]$. We assume that the programming language chosen for the implementation has vectorial operation capabilities (*e.g.* Matlab, R, Python with Numpy...), so that every operator $\odot$ between two arrays (typed in bold font) must be interpreted as a pointwise operator: $\mathbf{Z} = \mathbf{X} \odot \mathbf{Y}$ is a $N$-dimensional array defined by $\mathbf{Z}[i] = \mathbf{X}[i] \odot \mathbf{Y}[i]$. Similarly, any scalar operation $\mathbf{Z} = \lambda \odot \mathbf{X}$ (with $\lambda \in \mathbb{R}$) maps $\mathbf{X}$ to a $N$-dimensional array defined by $\mathbf{Z}[i] = \lambda \odot \mathbf{X}[i]$. Furthermore, for any pair of integers $i \leqslant j$, $\mathbf{X}[i:j]$ is the sub-array of $\mathbf{X}$ from the $i$-th to the $j$-th elements, and $i:j$ is a shortcut notation denoting a vector of consecutive integers ranging from $i$ to $j$. Eventually, & is the vector concatenation operator. If the algorithm is implemented with any other language, it is straightforward to implement these operations with basic loop instructions.

**Algorithm 1** Computation of the radial signature of a polygon

**Require:** Polygon vertex coordinates $\mathbf{X}$, $\mathbf{Y}$, an integer $n$ and an optional parameter $f$ (default value 10)
**Ensure:** A $n$-dimensional arrays $\mathbf{R}$

$N \leftarrow \text{LENGTH}(\mathbf{X})$
$\mathbf{X} \leftarrow \mathbf{X} - \text{MEAN}(\mathbf{X})$ `# To avoid numerical errors`
$\mathbf{Y} \leftarrow \mathbf{Y} - \text{MEAN}(\mathbf{Y})$ `# To avoid numerical errors`

`# Preparation`
$\mathbf{X_1} \leftarrow \mathbf{X}[1:(N-1)];\ \mathbf{X_2} \leftarrow \mathbf{X}[2:N];\quad \mathbf{Y_1} \leftarrow \mathbf{Y}[1:(N-1)];\ \mathbf{Y_2} \leftarrow \mathbf{Y}[2:N]$

`# Centroid computation`
$A \leftarrow 3 * \text{SUM}\big(\mathbf{X_1} * \mathbf{Y_2} - \mathbf{Y_1} * \mathbf{X_2}\big)$
$xc \leftarrow \text{SUM}\big((\mathbf{X_1} + \mathbf{X_2}) * (\mathbf{X_1} * \mathbf{Y_2} - \mathbf{Y_1} * \mathbf{X_2})\big)/A$
$yc \leftarrow \text{SUM}\big((\mathbf{Y_1} + \mathbf{Y_2}) * (\mathbf{X_1} * \mathbf{Y_2} - \mathbf{Y_1} * \mathbf{X_2})\big)/A$

`# Curvilinear abscissa preparation`
$\mathbf{X_s} \leftarrow \varnothing;\ \mathbf{Y_s} \leftarrow \varnothing;\ \mathbf{T} = (0:f)/f$
**for** $i = 0$ to $N-1$ **do**
  $\mathbf{X_s} \leftarrow \mathbf{X_s}\ \ \&\ \ (1-\mathbf{T}) * \mathbf{X}[i] + \mathbf{T} * \mathbf{X}[i+1]$
  $\mathbf{Y_s} \leftarrow \mathbf{Y_s}\ \ \&\ \ (1-\mathbf{T}) * \mathbf{Y}[i] + \mathbf{T} * \mathbf{Y}[i+1]$
**end for**

`# Radial distances computation`
$\mathbf{S} \leftarrow \text{CUMSUM}\big(\text{SQRT}\big((\mathbf{X_s}[2:n] - \mathbf{X_s}[1:(n-1)])^2 + (\mathbf{Y_s}[2:n] - \mathbf{Y_s}[1:(n-1)])^2\big)\big)$
$\mathbf{R} \leftarrow \text{SQRT}\big((xc - \mathbf{X_s})^2 + (yc - \mathbf{Y_s})^2\big)$
$\mathbf{R} \leftarrow \text{INTERP}(\mathbf{S}, \mathbf{R}, \mathbf{T})$
**return** $n * \mathbf{R}/\text{SUM}(\mathbf{R} * \mathbf{R})$

CUMSUM is a module taking as input a vector $\mathbf{X}$ and computing the cumulative sum of $\mathbf{X}$. For example, if $\mathbf{X} = [1, 2, 3, 4]$, then $\text{CUMSUM}(\mathbf{X}) = [1, 3, 6, 10]$. For 3 vectors $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{X}'$, the instruction $\text{INTERP}(\mathbf{X}, \mathbf{Y}, \mathbf{X}')$ performs linear interpolation of the

**Algorithm 2** Computation of the *radial distance* between two signatures

**Require:** A pair of $n$-dimensional floating point arrays $\mathbf{R_1}$ and $\mathbf{R_2}$
**Ensure:** A scalar positive float number representing the radial distance

1: $n \leftarrow \text{LENGTH}(\mathbf{R_1})$; $dmin \leftarrow$ `1e300`
2: **for** $k = 0$ to $n$ **do**
3: $\quad d \leftarrow \text{SUM}\big((\mathbf{R_1} - \text{SHIFT}(\mathbf{R_2}, k))^2\big)$
4: $\quad$ **if** $d < dmin$ **then**
5: $\quad\quad dmin \leftarrow d$
6: $\quad$ **end if**
7: **end for**
8: **return** $\text{SQRT}(dmin/n)$

function $\mathbf{Y} = f(\mathbf{X})$ to evaluate it along a new vector of sample points $\mathbf{X}'$.

The output of the algorithm 1 is an $n$-dimensional array $\mathbf{R}$ (where $n$ is the required size) with $\mathbf{R}[i]$ containing the value of the radial signature at the (normalized) curvilinear abscissa $i/n$.

Having the radial signatures of two polygons computed, the *radial distance* is evaluated with the following algorithm (2). We assume that SHIFT is a pre-existing module performing the circular shift of a vector. For example, if $\mathbf{X} = [1, 2, 3, 4, 5]$, then $\text{SHIFT}(\mathbf{X}, 2) = [4, 5, 1, 2, 3]$. The minimal distance is initialized on line 1 with an arbitrary large value (`1e300`), and then subsequently updated when better shifts are found.

### 4.2. *Acceleration of the computation using the analytical formula*

We consider two consecutive vertices of the polygon $\gamma$, $A$ and $B$. For vertex $A$, $s = s_A$ and $\rho = \rho_A$, and for vertex $B$, $s = s_B > s_A$ and $\rho = \rho_B$. We note $\Delta s = s_B - s_A$. $C(\gamma)$ is the center of mass of the polygon $\gamma$.

First, we suppose $\rho_B \geqslant \rho_A$. We note $H$ the projection of $C(\gamma)$ on $(AB)$, $h = C(\gamma)H$ and $l = AH$. Then, two cases can be identified:

**First case :** $\rho_B^2 > \rho_A^2 + \Delta s^2$.

In this case, the angle $\widehat{C(\gamma)AB}$ is greater than $\frac{\pi}{2}$ and $H$ is on the half-line $(AB]\setminus[AB]$. Then

$$\rho_A^2 = h^2 + l^2 \qquad \text{and} \qquad \rho_B^2 = h^2 + (l + \Delta s)^2,$$

which leads to

$$l = \frac{\rho_B^2 - \rho_A^2 - \Delta s^2}{2\Delta s} \qquad \text{and} \qquad h^2 = \rho_A^2 - \left(\frac{\rho_B^2 - \rho_A^2 - \Delta s^2}{2\Delta s}\right)^2.$$

Finally, the analytical formula, for $u \in [0, \Delta s]$ is:,

$$\rho(s_A + u) = h^2 + (l + u)^2 = \rho_A^2 - \left(\frac{\rho_B^2 - \rho_A^2 - \Delta s^2}{2\Delta s}\right)^2 + \left(\frac{\rho_B^2 - \rho_A^2 - \Delta s^2}{2\Delta s} + u\right)^2.$$

**Second case :** $\rho_B^2 \leqslant \rho_A^2 + \Delta s^2$.

In this case, the angle $\widehat{C(\gamma)AB}$ is smaller than (or equal to) $\frac{\pi}{2}$ and $H$ is on $[AB]$. Then

$$\rho_A^2 = h^2 + l^2 \qquad \text{and} \qquad \rho_B^2 = h^2 + (\Delta s - l)^2,$$

which leads to

$$l = \frac{\rho_A^2 + \Delta s^2 - \rho_B^2}{2\Delta s} \qquad \text{and} \qquad h^2 = \rho_A^2 - \left(\frac{\rho_A^2 + \Delta s^2 - \rho_B^2}{2\Delta s}\right)^2.$$

Finally, the analytical formula for $u \in [0, \Delta s]$ is:

$$\rho(s_A + u) = h^2 + (l - u)^2 = \rho_A^2 - \left(\frac{\rho_A^2 + \Delta s^2 - \rho_B^2}{2\Delta s}\right)^2 + \left(\frac{\rho_A^2 + \Delta s^2 - \rho_B^2}{2\Delta s} - u\right)^2.$$

To be complete, when $\rho_A \geqslant \rho_B$, then if $\rho_A^2 > \rho_B^2 + \Delta s^2$, then:

$$\rho(s_A + u) = \rho_B^2 - \left(\frac{\rho_A^2 - \rho_B^2 - \Delta s^2}{2\Delta s}\right)^2 + \left(\frac{\rho_A^2 - \rho_B^2 - \Delta s^2}{2\Delta s} + (\Delta s - u)\right)^2,$$

and if $\rho_A^2 \leqslant \rho_B^2 + \Delta s^2$, then:

$$\rho(s_A + u) = \rho_B^2 - \left(\frac{\rho_B^2 + \Delta s^2 - \rho_A^2}{2\Delta s}\right)^2 + \left(\frac{\rho_B^2 + \Delta s^2 - \rho_A^2}{2\Delta s} - (\Delta s - u)\right)^2.$$

The analytical interpolation formula above may be implemented in the algorithm 1, line 23, instead of pointwise distance computations.

Computation of the analytical radial signature provides three significant assets. First, it enables to compute the radial signature virtually with an infinite resolution. Second, it provides a nearly closed-form expression for the radial signature, which may ease subsequent theoretical development, in particular regarding the modeling of data positional uncertainty propagation on the *radial distance*. And third, it has been shown with experimentation that computing radial signature based on a regular sampling of its analytical form is faster than oversampling the original polygon before computing the radial signature, as illustrated on Figure 8. The rationale behind this observation, is that oversampling is a systematic and straightforward process which does not take too much computational burden, while radial signature evaluation requires to evaluate expensive distances. Therefore, it seems profitable to start with the latter on a sparse polygon, and then to interpolate the result with the analytical form, while on the contrary, starting with polygon oversampling compels to estimate

the radial signature on a large number of points. The experimentation revealed that, for a reasonable amount of oversampling (about 1 point every meter on the original polygon) the process is sped up by a factor of 2.5 with the analytical form. Other tests on our dataset of buildings have shown that to obtain a precision of 1% on the radial signature, one needs to take an oversampling factor of 11 or higher, which corresponds to a point every few meters, for typical building edge lengths.

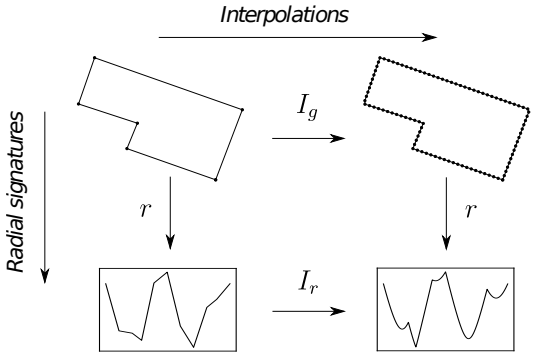


**Figure 8.** Two possible computation paths to estimate an accurate radial signature (bottom right) from a polygon (top left): 1. with an interpolation $I_g$ in the geographical space, followed by radial signature computation $r$, or 2. more efficiently, first radial signature computation $r$ on the raw polygon, followed by an interpolation $I_r$ in the radial signature space, using the analytical form.

### 4.3. *Computation with Fast Fourier Transform*

For numerical purpose, both $\gamma$ and $\gamma'$ have to be discretized into $n$ points. In general, $n$ corresponds to the number of vertices of the original polygons, but they may as well have been oversampled during the radial signature computation step. In the subsequent time complexity analysis, for the general case where $\gamma$ and $\gamma'$ have different size $m \neq m'$, we denote $n = \max(m, m')$.

The straightforward algorithm for computing the *radial distance* (3.3) implies to test a set of $n$ possible translations $\tau_1$, $\tau_2$,... $\tau_n$ in $[0, 1]$, and for each $\tau_i$, to evaluate a discrete integral of $n$ terms, therefore, the straightforward computation of the expression in Definition 3.3 requires $\Theta(n^2)$ time complexity. Hereafter, we propose a fast algorithm, based on the Fast Fourier Transforms (FFT). The term to minimize over $\tau$ in Definition 3.3 can be written in the following way :

$$\begin{aligned}\int_0^1 \big(r_\gamma(t) - r_{\gamma'}(t+\tau)\big)^2 dt &= \int_0^1 r_\gamma(t)^2 dt + \int_0^1 r_{\gamma'}(t+\tau)^2 dt - 2\int_0^1 r_\gamma(t) r_{\gamma'}(t+\tau)^2 \\ &= \|r_\gamma\|^2_{L^2} + \|r_{\gamma'}\|^2_{L^2} - 2\int_0^1 r_\gamma(t) r_{\gamma'}(t+\tau)\end{aligned}$$

with the variable substitution $t \mapsto t+\tau$ giving: $\int_0^1 r_{\gamma'}(t+\tau)^2 dt = \int_0^1 r_{\gamma'}(t)^2 dt = \|r_{\gamma'}\|^2_{L^2}$. The first two terms of the above quantity do not depend on $\tau$, and the optimal shift

may be computed by evaluating :

$$\bar{d}_r(\gamma,\gamma') = \max_{\tau\in[0,1]} \int_0^1 r_\gamma(t) r_{\gamma'}(t+\tau)dt = \max_{\tau\in[0,1]} \int_0^1 r_\gamma(t) r_{\gamma'}(t-\tau)dt = \max_{\tau\in[0,1]} \big(r_\gamma * r_{\gamma'}\big)(\tau)$$

where $*$ denotes the convolution operator on $L^2([0,1])$.

However, Fourier Transform operator $\mathcal{F}$ maps convolution to pointwise multiplication in temporal and spectral spaces, respectively. Then, denoting $\mathcal{F}^{-1}$ the inverse Fourier Transform, the formula above can be rewritten as:

$$\bar{d}_r(\gamma,\gamma') = \max_{\tau\in[0,1]} \Big(\mathcal{F}^{-1}\big[\mathcal{F}[r_\gamma]\big] * \mathcal{F}^{-1}\big[\mathcal{F}[r_{\gamma'}]\big]\Big)(\tau) = \max_{\tau\in[0,1]} \Big(\mathcal{F}^{-1}\big[\mathcal{F}[r_\gamma]\mathcal{F}[r_{\gamma'}]\big]\Big)(\tau)$$

The computations of $r_\gamma$ and $r_{\gamma'}$ requires $\Theta(n)$ operations (computing the centroid and then the distance to the centroid for each of the $n$ discretized points). Using FFT algorithm, $R_\gamma = \mathcal{F}[r_\gamma]$ and $R_{\gamma'} = \mathcal{F}[r_{\gamma'}]$ can be computed in $\Theta(n \log n)$, and the pointwise multiplication $R_\gamma R_{\gamma'}$ requires $\Theta(n)$ floating point operations. Eventually, the inverse FFT takes $\Theta(n \log n)$ operations, and an additional $\Theta(n)$ operations are required to find the optimal shift $\tau$. Considering all the operations results in a $\Theta(n \log n)$ algorithm for computing the *radial distance*. All the implementations would take place in Algorithm 2, between lines 2 and 6.

The FFT approach has been implemented and tested on a set of polygons. Experimental results revealed that it clearly outperforms the straightforward approach. With the standard algorithm, and for simple polygons (up to 20 vertices), about 100 *radial distances* may be computed every second. However, in the context of geographic data matching, the number of comparisons to perform is generally one order of magnitude greater than the number of polygons in the databases, since it is often required to compare each polygon in a database, with a certain number (often up to 10) candidates in the second database. Hence, with this algorithm, matching two datasets of 36 000 buildings requires about 60 minutes. This estimation is pessimistic, since experimentation revealed that taking advantage of the vectorization capabilities of some of the programming languages (in particular R or Matlab), it is easy to achieve up to 300 distance computations per second, but this is still a fairly long amount of time for the evaluation of a single criteria in the matching process.

However, the proposed FFT-based approach enables to reach a computation speed of 500 polygons per second, even for moderately complex shapes (up to 500 vertices). Even with polygons containing a million of vertices, the *radial distance* computation may be performed in less than one second, while its evaluation would be impossible in a reasonable amount of time with the standard algorithm.

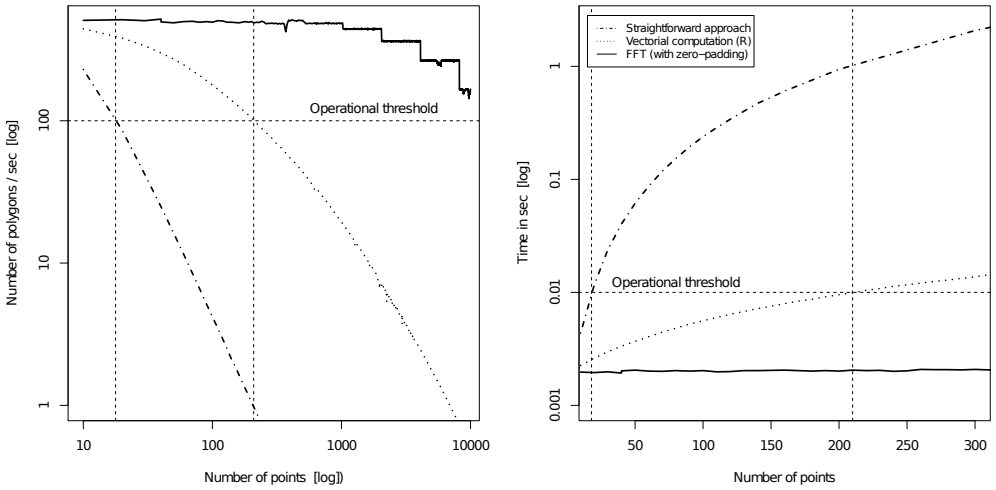


**Figure 9.** Comparison of computational performances of 3 algorithms: FFT-based approach (plain line) and straightforward approach with (dashed line) and without (dotted line) vectorial computation capabilities versus number of points in the radial signatures. Left: number of polygon computed (in log scales). Right: computation time in seconds (in log scale). The horizontal line *Operational threshold* depictcs the minimum requirements to be able to match two datasets of 36 000 buildings in less than 1 hour, with an average of 10 homologous candidates per polygon.

It is interesting to note that on Figure 9 (left), the straightforward approach shows a linear pattern, which was expected since its quadratic complexity turns out to be linear in log-log scale. Besides, the discontinuity of the FFT-based approach is due to the *zero-padding process*: when the number $n$ of points in the signature is not a power of two, the corresponding building is linearly interpolated to reach $2^{\lceil \log_2 n \rceil}$ points.

## 5. Robustness of the *radial distance*

In this section, we evaluate how the *radial distance* fluctuates when one point of the polygon is moved, by computing an analytical majoration of the variation of the *radial distance* at the first order.

We consider a polygon $\mathbf{P} = P_1 \dots P_N$, and compute the variation of the *radial distance* when one point $P_i$ is modified into $\widehat{P_i} = P_i + \delta P_i$. In this section, the $\widehat{\cdot}$ symbol will designate quantities associated with the modified polygon, and $\delta\cdot$ the difference between the modified quantity and the initial quantity (for example $\delta L = \widehat{L} - L$, $L$ being the perimeter of the polygon).

We note $P^-$ and $P^+$ the neighbouring points of $P = P_i$ in $\mathbf{P}$ and $\mathbf{T}$ the triangle $P^-PP^+$. For simplicity, we suppose that $\mathbf{T} \subset \mathbf{P}$ (when $\mathbf{P}$ and $\mathbf{T}$ are considered as surfaces). We note $l_1 = P^-P$, $l_2 = PP^+$, $l_3 = P^-P^+$, $\alpha$ the angle $(PP^-, PP^+)$, $\mathcal{A}$ the area of $\mathbf{P}$, $\mathcal{A}_\mathbf{T}$ the area of $\mathbf{T}$, and finally $G$, $G_{\mathbf{P}\setminus\mathbf{T}}$ and $G_\mathbf{T}$ the centers of mass of $\mathbf{P}$, $\mathbf{P}\setminus\mathbf{T}$ and $\mathbf{T}$, respectively.

Several factors contribute to the variation of the *radial distance*: the displacement of $P$ changes the distance between $G$ and $P$, but also modifies $G$ (changing the radial signature at all points), $L$ (which changes the perimeter normalization factor for the

linear abscissa), and ultimately the $L^2$ normalization factor.

To quantify these variations, we note $\delta = \|\delta P\|$, $\varepsilon_1 = |\delta P \cdot \nu_1|$, $\varepsilon_2 = |\delta P \cdot \nu_2|$, where $\nu_1$ is a unitary vector normal to $(P^-P^+)$, $\nu_2$ is a unitary vector normal to the ellipse $\mathcal{E}(P^-, P^+; P)$ at point $P$ ($\mathcal{E}(P^-, P^+; P)$ notes the ellipse of focal points $P^-$ and $P^+$ through $P$). Note that the vectors $\nu_1$ and $\nu_2$ are not related in general. Then we have the following majoration for the *radial distance*:

$$d_r(\mathbf{P}, \widehat{\mathbf{P}})^2 \leqslant \frac{1}{R(\mathbf{P})^2} \left( C_1 \delta^2 \left( \frac{l_1 + l_2}{L} \right)^2 + C_2 \varepsilon_2^2 \sin^2 \left( \frac{\pi - \alpha}{2} \right) + C_3 \left( \frac{\mathcal{A}_\mathbf{T}}{\mathcal{A}} \right)^2 \delta^2 + C_4 \left( \frac{l_3}{\mathcal{A}} \right)^2 \varepsilon_1^2 \left( G G_\mathbf{T} \right)^2 \right)$$

with $C_1$, $C_2$, $C_3$ and $C_4$ constants independent of $\mathbf{P}$.

The displacement of a point modifies the radial signature through different effects, and in a first order approach, these effects generate distinct terms in the majoring quantity. The first term accounts for the displacement of the points of the segments $[P^-P]$ and $[PP^+]$. The second terms accounts for the change of the normalized curvilinear abscissa of all points of $\mathbf{P}$ when the perimeter changes. The last two terms come from the displacement of the mass center of the polygon, the third quantifying how the displacement of the mass center of $\mathbf{T}$ affects the position of the mass center of $\mathbf{P}$, the fourth how the variation of the area of $\mathbf{T}$ affects the center of mass of $\mathbf{P}$ (only the component $\eta$ of $\delta P$ intervenes in this last term because perturbations collinear to $P^-P^+$ do not change the area of $\mathbf{T}$).

$L^2$ normalization generates a multiplicative factor (of at most 2) for the majoration. The radial signatures of the original polygon and of the modified polygon may be shifted for the computation of the *radial distance*, but this process can only lead to a smaller difference and does not affect the majoration. The interactions between these different effects may only generate terms of order 2 or more in $\delta P$, so we do not take them into account in the majoration at the first order.

We note that the modifications of the perimeter and of the center of mass of the polygon generate changes that are not restricted to the immediate neighborhood of $P$. For the perturbation of a single point $P$, the radial signature of a polygon is more sensitive to these global changes if:

- $\alpha$ is small, that is, the angle at $P$ is very acute (variation of the perimeter),
- the area of $\mathbf{T}$ is big (displacement of the center of mass of $\mathbf{T}$),
- the distance between the centers of mass of $\mathbf{T}$ and $\mathbf{P}$ is big, which is often the case when $P$ is distant from $G$ (modification of the area of $\mathbf{T}$).

Some of these effects (variations of $L$ or $\mathcal{A}_\mathbf{T}$) are only triggered by one of the components of $\delta P$. The perturbation of $P$ also creates a localized effect on segments $[P^-P]$ and $[PP^+]$, and this effect is more important when $l_1 + l_2$ is big, that is, when $P$ is far from its neighbors. These sensibility factors are not independent, for example, a small value of $\alpha$ can be associated with a point $P$ far from $G$,

but with a small area for $\mathbf{T}$. Big values of $l_1+l_2$ can generate big values of the area of $\mathbf{T}$.

In general, we conclude that the radial signature is particularly sensitive to perturbations for polygons with eccentric points, very acute angles, and segments that are long (relative to the perimeter).

## 6. Experiments on real data

The aim of this section is to report and discuss the results of a set of experiments conducted to grasp the behavior of the *radial distance* on real polygon shapes representing buildings. To do so, 29 000 buildings on Val-de-Marne area (France) are extracted from both OpenStreetMaps (OSM) and the authoritative topographic database (BD TOPO) produced by the French National Mapping Agency. The two datasets have been matched by using the data matching algorithm based on belief function theory described in Maidaneh Abdi *et al.* (2020). Among the matched polygons, 391 of them, located on three different representative districts, are sampled based on the recommendations proposed by Foody (2009) and manually validated (incorrect matching links have been discarded, so that subsequent experiments are conducted on a reliable dataset of homologous polygons).

### 6.1. *Experimentations*

We defined four experiments, each with a specific goal. Experiment 1 aims at finding a rough separation criterion between homologous and non-homologous pairs of buildings based on the respective distributions of *radial distance* values. Experiment 2 aims at measuring the sensibility of *radial distance* output on different types and amplitudes of noise on data. It is a simulation based experiment that addresses only perturbations due to survey noise and cartographic generalization. As it would require a very specific and *ad hoc* statistical modeling, the additional perturbations caused by the respective databases *nominal terrain* descriptions are not handled in this experimentation (see pipeline depicted on Figure 4). The results of this experiment contain multiple practical abacuses to decide whether a pair of candidate buildings are likely to be homologous, given their mutual *radial distance* and the level of noises contaminating their respective datasets. Experiment 3 seeks to describe how the respective noises perturbing two datasets are adding up together on *radial distances*. Eventually, we conclude this section by demonstrating that simulations can also be used *online* to determine whether two polygons are homologous.

In this section, we denote $H_0$ the null hypothesis, stating that two polygons are actually homologous. From a statistical perspective, the goal is to reject $H_0$. If we compute a given *radial distance* between two polygons, and the probability that two homologous polygons (with stochastic noise) are separated by a given *radial distance* is lower than a p-value threshold (5%), $H_0$ is rejected (or equivalently, $H_1$ is selected) and we conclude that the two polygons are not homologous. Conversely, if the obtained p-value is greater than the threshold, $H_0$ cannot be rejected (we may abusively say that the two polygons are homologous; more rigorously, we should conclude that there is no statistical evidence that they are not homologous).

### E1. Distributions of *radial distance* values

We evaluated the *radial distances* of the 391 homologous building pairs and the estimated distributions of computed values are provided in Figure 10. The same process is applied to 391 randomly sampled pairs of non-homologous buildings. However, it may be argued that for practical use, polygon data matching if usually performed with neighbor candidates, which may statistically be more similar than merely random samples of building pairs on the full dataset. Therefore, we computed a third distribution, of *radial distance* values of non-homologous buildings - yet still neighbor (*i.e.* closer than 50 m) - buildings. It turned out that, surprisingly, there is no statistically significant difference between the distributions of randomly sampled pairs and pairs of neighbor buildings. This may be explained by the fact that there is no local shape homogeneity.

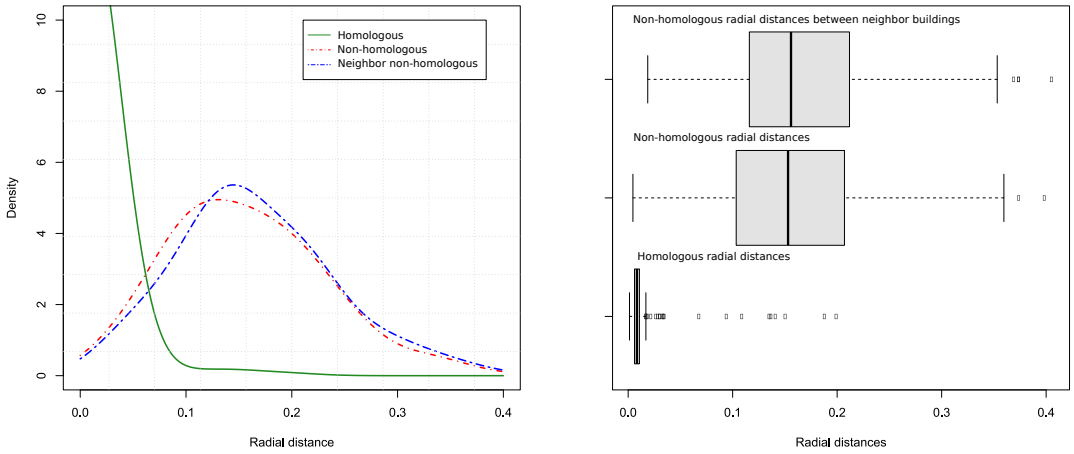


**Figure 10.** Distributions of *radial distance* values for (a) homologous buildings (green plain line - bottom boxplot), (b) non-homologous and random pairs of buildings (red dashed line - middle boxplot) and (c) non-homologous and random pairs among neighbors (closer than 50 m) buildings (blue dotted line - top boxplot).

From the results depicted on Figure 10, it seems clear that when the *radial distance* value is above around 0.07, the likelihood turns in favor of non-homologous buildings hypothesis. It also reveals that, despite its theoretical supremum value of $\sqrt{2}$ (as demonstrated in Proposition 3.8), in facts, the *radial distance* hardly gets above 0.4 at the maximum for very dissimilar buildings. The example exhibited in Proposition 3.8 is not realistic in practice, and it is important to keep this in mind when setting a *radial distance* threshold in any practical implementations.

Besides, these simple results offer an opportunity for solving the problem of deciding whether two buildings are homologous, with Bayesian theory:

$$\mathbb{P}(H_0|R=r) = \frac{\pi(R=r|H_0)\mathbb{P}(H_0)}{\sum_{i\in\{0,1\}}\pi(R=r|H_i)\mathbb{P}(H_i)} = \frac{\pi(R=r|H_0)\mathbb{P}(H_0)}{\pi(R=r|H_0)\mathbb{P}(H_0)+\pi(R=r|H_1)\mathbb{P}(H_1)}$$

where $\pi(R|H_i)$ stands for the conditional density of the continuous variable $R$ given that hypothesis $H_i$ $(i = 0, 1)$ is true.

For example, applying this formula for a measured *radial distance* of $r = 0.05$, on two datasets with 90% completion and with a search of 10 candidates for each feature to match, there is a prior $\mathbb{P}(H_0) = 0.9 \times (1/10) = 0.09$ and it is found that $\mathbb{P}(H_0|R = r) = 0.149$, hence 14.9% probability that the two buildings are homologous. Increasing the number of tested candidates or matching the building with a more incomplete dataset would decrease this probability value.

### E2. Robustness of *radial distance* on survey and generalization noises

The results of experiment 1 are convenient to give a matching criteria threshold on any *radial distance* computed between polygons extracted respectively from OSM and BD TOPO (or any other two databases having similar noise properties). However, this result cannot be used as such for all databases.

In this second experiment, we use noise simulation to assess the sensibility of *radial distance* evaluations on dataset noises. The motivation behind this, is to be able to determine, for a given type and level of noise, the practical[1] maximal value of *radial distance* separating two homologous buildings.

To generate realistic noise, we used an approach described in Ripley (2009): with a random generator, we sampled $n$ *i.i.d.* unit-variance and zero-mean gaussian values, compiled in a vector $\mathbf{x}$. It can easily be shown that, for any positive-definite matrix $\mathbf{\Sigma} \in \mathbb{R}^{n \times n}$, the random vector $\mathbf{y} = \mathbf{A}\mathbf{x}$ where $\mathbf{A}$ is a Cholesky factor of $\mathbf{\Sigma}$, is a realization of a correlated random vector $\mathbf{Y}$ having covariance matrix $\mathbf{\Sigma}$. We used this process separately on both axes of the euclidian plane, to add a correlated noise to any building polygon shapes. The covariance matrix $\mathbf{\Sigma}$ is formed with a (stationary) covariance kernel with three parameters:

- The **type of kernel**: exponential, gaussian, and triangular models are used.
- The **amplitude** of noise: between 0 and 5 meters, as it is quite uncommon to find building databases with more than 5 m error amplitude. If necessary, the output tables could be extended to handle large errors.
- The correlation **scope** of the noise which roughly speaking, describes how far apart two errors would remain correlated (in both amplitude and direction): between 1 m (white noise) and 1000 m (global translation).

On the graphics below, correlation scope is expressed relatively to the average diameter of buildings in the database ($\sim$ 10 m), meaning that, in logarithmic scale a value of -1 corresponds to 1 m ($10^{-1}\times$ diameter) while a value of 2 corresponds to

[1]Here, by *maximal practical* value, we may refer to the 95th percentile value, which would directly correspond to the 5% risk rejection threshold for $H_0$.

1000 m ($10^2\times$ diameter).

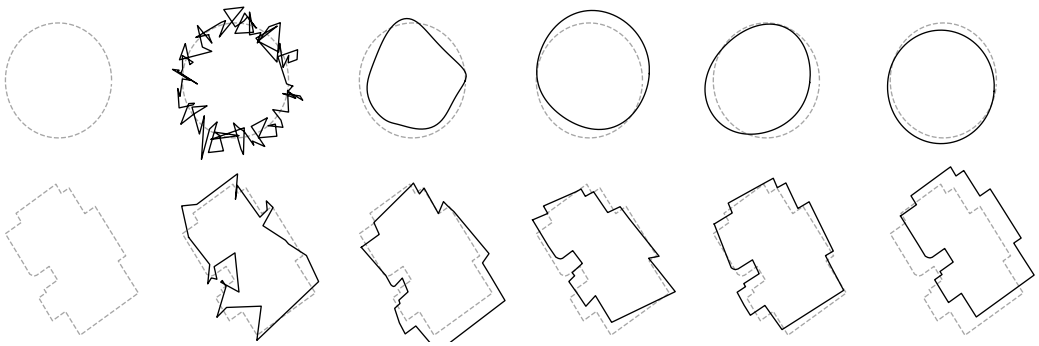

**Figure 11.** Illustration of the noise model on a 30-meter-radius continuous circle (top) and on a piecewise linear building (bottom). Covariance scope is increasing from 0.1 (white noise process; left) to 500 m (global translation; right), for a same amplitude level of 5 meters. Original shapes are depicted with dashed lines.

The noise amplitude levels were sampled continuously between 0 and 5 m, and then divided into 19 groups for the matrix representation below. On the other hand correlation scopes were discretized regularly on 13 levels between 1 and 1000 m along a logarithmic scale. In total, 100 000 simulations have been performed for each covariance model, resulting in at least 300 samplings in each of the $19 \times 13$ cells of the experiment design, ensuring statistical robustness of the computed values (figure 13).

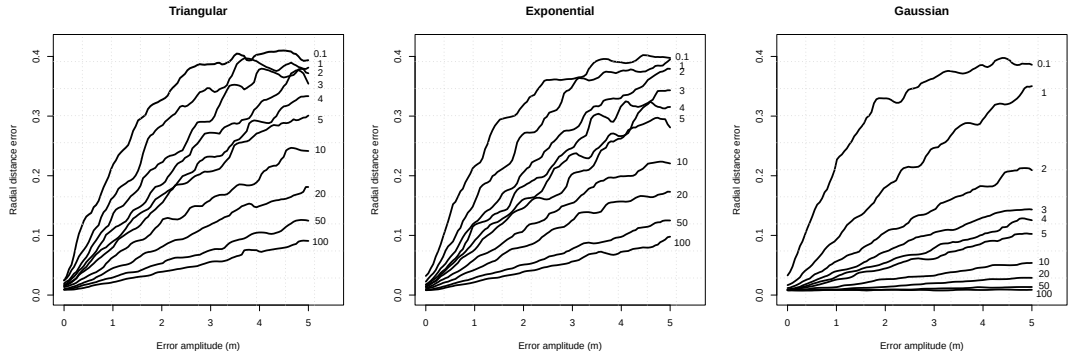


**Figure 12.** Quantile regression of 95th percentile of *radial distance* evaluation error versus noise amplitude level for different correlation scopes (indicated on the right of plots) and different covariance models (from left to right: triangular, exponential and gaussian).

Triangular and exponential kernels give very similar results while gaussian kernel outputs are slightly different. It is interesting to note on Figure 12 that for all correlation scopes/models, the impact of the noise amplitude on the *radial distance* error seems to be approximately linear, as long as the noise level is sufficiently small, then a plateau is reached at a value depending essentially on the covariance scope, but also on the covariance model: for a same scope, the plateaus of the Gaussian model are clearly lower than those of the exponential and triangular kernels.

Figure 13 can be readily used as a decision table as follows: suppose that two potentially homologous buildings are compared, separated by a *radial distance* $r = 0.15$.

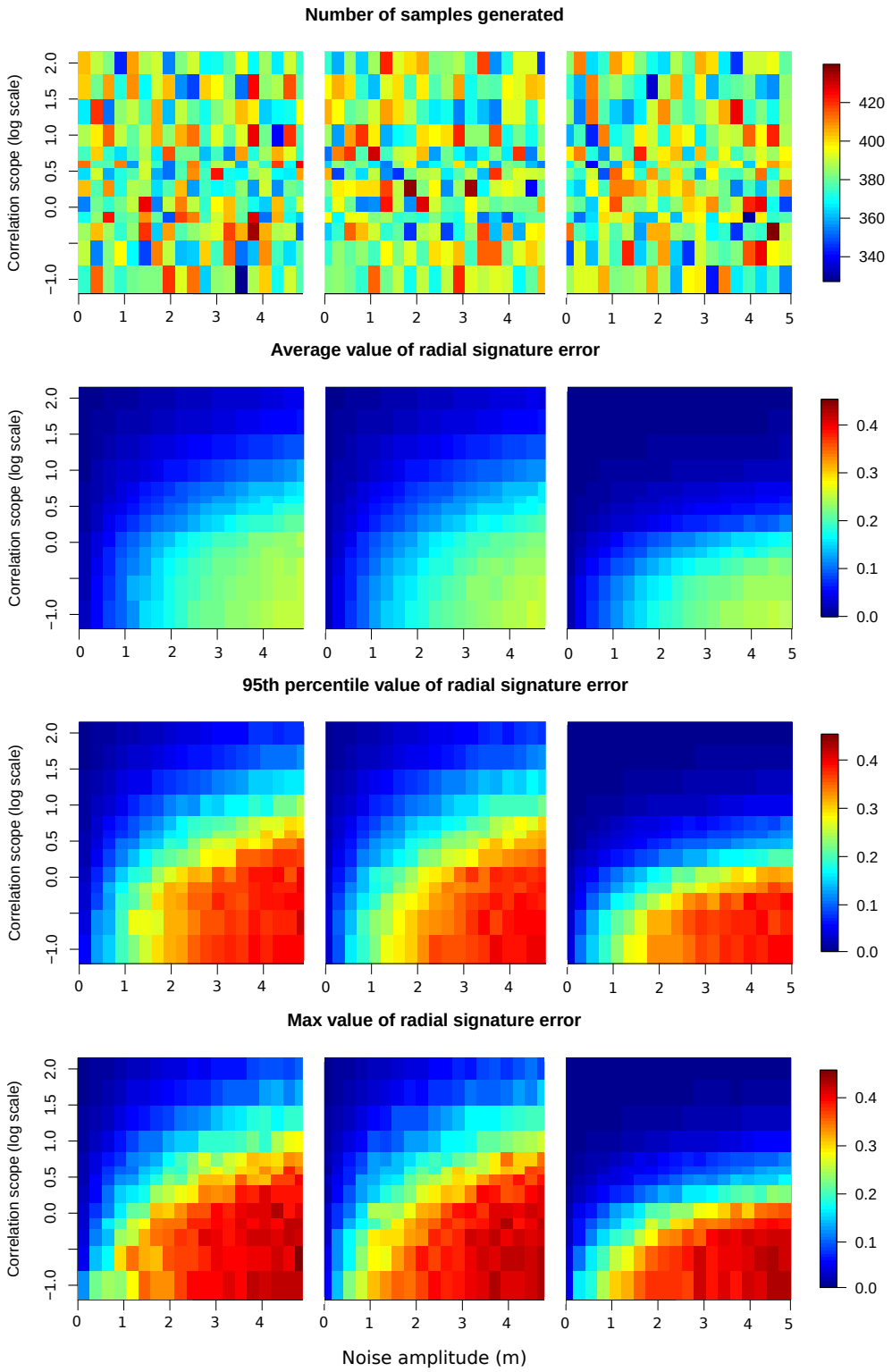


**Figure 13.** From top to bottom: (1) number of simulations, (2) mean *radial distance* error, (3) $95^{th}$ percentile of *radial distance* error and (4) maximal value of *radial distance* error. Each cell value is referring to a noise amplitude (in $x$-axis) and a noise covariance scope (in $y$-axis). The experimentation has been performed for 3 different noise covariance models (from left to right: triangular, exponential and gaussian).

Further, let us suppose that the homologous candidate stems from a database known to be contaminated by 2-meter amplitude with 100 meter-correlation noise (we assume triangular or exponential covariance model). Then, in the 95th percentile table ($x = 2\ m$ and $y = 1$), a value around 0.12 can be read. Thus, null hypothesis $H_0$ can be rejected and state that these two buildings are unlikely to be homologous (*i.e.* given the noise specification, an evaluated *radial distance* of 0.15 is too far from 0 to be explainable by noise only).

We conducted a similar experimentation to evaluate the impact of cartographic generalization on *radial distance*. Generalization is the process of modifying geometric features to make them more easily readable from the map user's perspective (Buttenfield 1991). It relies on a trade off between geometric accuracy and simplicity. In this experiment, we assess the impact of two representative algorithms:

- A polyline simplification algorithm, which seeks to reduce the number of points while preserving most of the original shape of the building. We used the popular algorithm of Douglas and Peucker (1973). Note that, for the sake of completeness, we also experimented with Visvalingam and Whyatt (1993) algorithm, but found no significant difference, hence only Douglas-Peucker results are reported below.

- A footprint squaring algorithm (illustrated on Figure 14), ensuring that nearly flat and right-angles are rounded exactly to 0 and 90 degrees, respectively. We used a least squares approach based on the one described by Lokhat and Touya (2016), except that constraints on 45 degrees angles and parallel walls are not enforced in our experiments.

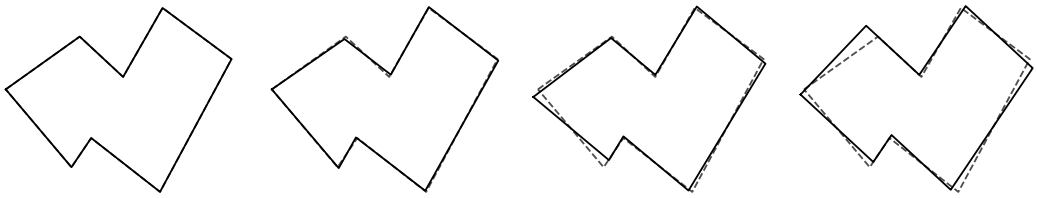

**Figure 14.** Illustration of building squaring for cartographic generalization, with a threshold tolerance angle (from left to right): 1°, 5°, 10° and 30°. On each picture, the original building is depicted with dashed lines (note that no modification has been undergone by the building with 1° threshold).

The generalization level is parametrized with a distance tolerance threshold (for Douglas-Peuker) and an angular tolerance threshold (for squaring). For each algorithm 10 000 simulations have been performed.

First, the results (Figure 15) show that Douglas-Peucker generalization process may potentially be significantly more destructive on the *radial distance* than survey acquisition noise, with a plateau value of its 95th percentile at 0.45 (which is even more than for white noise perturbation). For a typical and realistic value of 2 m generalization level, the average radial error is 0.05, with a 95th percentile error at 0.2. The same value (2 m) taken as a standard deviation of a moderately correlated noise, results in a 95th percentile error equal to 0.1 on *radial distance*, hence suggesting

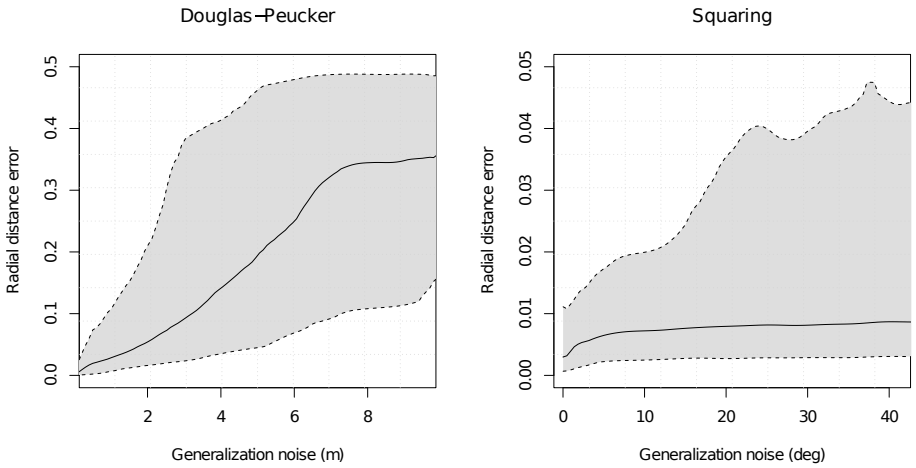


**Figure 15.** Median (plain line), $5^{th}$ and $95^{th}$ percentiles (dashed line) estimations of *radial distance* propagated error versus generalization noise (in meters/degrees) for 2 different generalization algorithms (left: Douglas-Peucker, and right: squaring).

that random perturbation of vertices is more conservative from the *radial distance* point of view, than the deterministic Douglas-Peucker algorithm. This may be easily explained by the fact that Douglas-Peucker (or any other simplification algorithm) may be likely to modify a building in a located and dissymetrical manner, which may right away displace significantly the polygon center of mass. Such a coordinated modification is unlikely to happen with a random perturbation.

However, things are surprisingly different with squaring algorithm, which seems to have very little impact on average. Even taking the 95th percentile results in very moderate changes in *radial distance* (0.05 at the maximum for a nearly 45 degrees angular tolerance threshold). This should be viewed in the light of Figure 10, where the distribution intersection between homologous and non-homologous pairs of buildings was occurring at 0.07. As a conclusion, it may be considered that squaring-type generalization operations can be neglected regarding their impact on *radial distance* evaluation, *i.e.* squaring modifications alone can hardly explain a significant value measured between two homologous buildings.

### E3. Additive properties of noise on *radial distances*

It is well known that when two independent noise contributions $\sigma_1$ and $\sigma_2$ add up, the amplitude of their resultant is a quadratic sum of the individual amplitudes: $\sigma^2 = \sigma_1^2 + \sigma_2^2$. In this experiment, we investigate whether this is true also for the impact of noise of *radial distances*.

Figure 16 depicts the result of 350 000 simulations, using a similar approach as described in the first part of experiment 2. For each simulation, a moderately correlated (50 m scope) noise is introduced independently on two versions of a same building. The *radial distance* between the two resulting polygons is evaluated, and results are aggregated in a matrix. It can be observed that the isolines of left matrix exhibit circular patterns, which confirms that noise on *radial distance* is adding

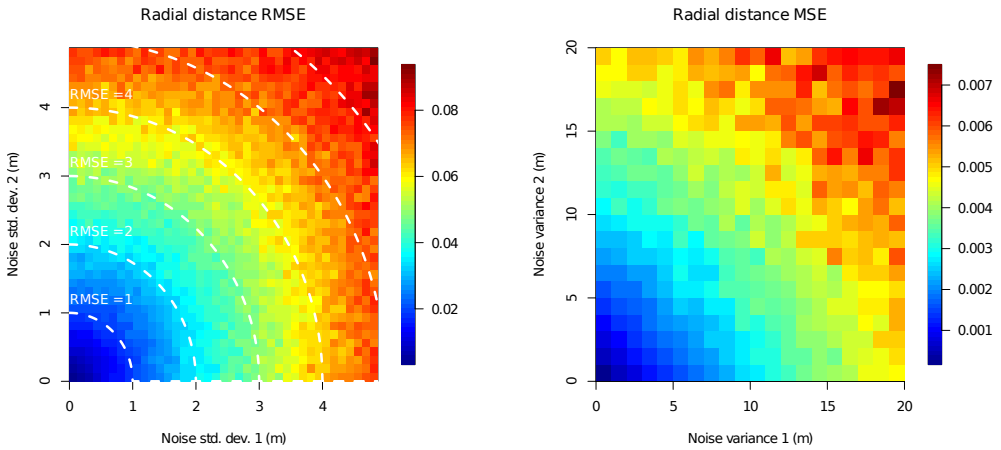


**Figure 16.** Error on *radial distance* evaluation versus noises on database 1 and database 2. Left: in terms of RMSE (Root Mean Square Error). Right: in terms of MSE (Mean Square Error).

quadratically. This is even more visible on the right graphic, depicted in terms of Mean Square Error (MSE), where isolines are a sequence of $-45°$ slope straight lines, as it could be expected.

This result is important, since it enables to disregard the uncertainty on one of the two databases involved in the matching process when its noise is significantly small in front of the noise level contaminating the other database. For example, let us consider two moderately correlated noise perturbating two databases $\mathcal{D}_1$ and $\mathcal{D}_2$, with respective amplitudes $\sigma_1 = 50$ cm and $\sigma_2 = 1$ m. Then the effects on the *radial distance* will be approximately the same than considering those of a perfect ideal database matched with another database of noise level : $\sqrt{0.5^2 + 1^2} = 1.12$ m, meaning that the noise on the database $\mathcal{D}_1$ is almost negligible.

### E4. Simulations for online polygon matching testing

In the first experiment, we extracted a criterion to decide whether two polygons are homologous, for the restricted case of OSM and BD TOPO databases. Experiment 2 was designed to address the case where databases have different kinds of noise perturbations. However, all the decision criteria compiled in output tables so far do not take into account the specific shapes of buildings, despite the fact that some polygon shapes may be particularly more sensitive, and then more likely to undergo large deviations on their *radial distance* evaluation. In particular, we showed in section 5 that angle values of the polygon may be an important factor explaining whether a polygon is sensitive or not.

In this last sub-section, we provide an experiment suggesting that online simulations may enable to accept or reject the null hypothesis with a higher confidence level.

We consider a particular pair of buildings $A$ and $B$, for which the *radial distance* has been evaluated, and we want to know if they are homologous. The idea involved here consists in simulating a number $n$ of perturbations of $A$ with the noise level of the database contaminating $B$. It results in $N$ noised versions $A_1$, $A_2$,... $A_n$ of $A$, each of them being a potential representation of the building in the second database.

The $n$ values of *radial distances* $d_i = d_r(A_i, B)$ are computed and this distribution of values is compared with the really observed *radial distance* $d = d_r(A, B)$. Then a simple estimation of the p-value in favor of null hypothesis $H_0$ may be calculated as the ratio of simulations $d_i$ being above $d$. $H_0$ is classically rejected if $p < 0.05$ and the two buildings are considered as non-homologous.

We estimated the covariance kernel between OSM and BD TOPO building polygons, by performing a pointwise matching with a Dynamic Time Warping approach (Müller 2007) on the 391 matched polygons. The resulting experimental kernel was then used to adjust a parametric mixture model (figure 17): a mixture of Gaussian (90 %) and triangular model (10 %), with respective scopes equal to 50 and 100 meters, and with a maximal standard deviation value of 1.52 meters (figure 17).

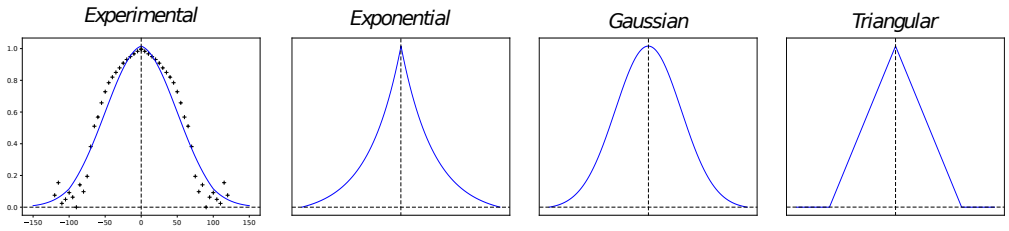


**Figure 17.** Left: the estimated covariance model between OSM and BD TOPO building polygons. Experimental kernel is depicted in black crosses, adjusted model in blue line. Right: the 3 main covariance kernels used as input in the mixture model (from left to right: Exponential, Gaussian and Triangular).

Examples of simulations are provided on Figure 18. In each case, we compare the *radial distances* between blue and red polygons, with the distributions of *radial distances* between blue and simulated gray curves. If the latter is a significant outlier, then $H_0$ is rejected and polygons are assumed to be non-homologous.

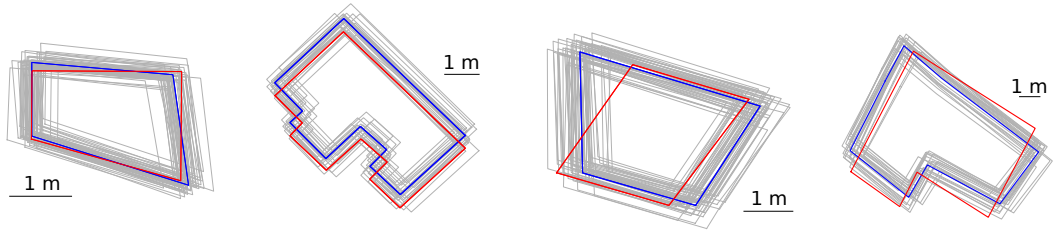


**Figure 18.** Four example of simulations. Blue: BD TOPO building. Red: OSM building. Gray: noised versions of BD TOPO building. The distribution of *radial distances* between blue and gray polygons is compared to the observed *radial distance* separating blue and red polygon. For the right two polygons, p-value was above 0.05 and $H_0$ was rejected.

We performed simulations for the 391 pairs of homologous buildings and for 391 randomly sampled pairs of buildings (see experiment 1). Figure 19 depicts the respective distributions of p-values. It can be seen clearly that the resolving power of this representation is much better than the one of the raw *radial distance* values (Figure 10). With a threshold set at a p-value of 0.05, only 1.7 % of homologous buildings are rejected, while 14 % of non-homologous buildings are considered as homologous. With a more symmetrical threshold of 0.2, the separation is even more

accurate, with about 3 % of homologous buildings rejection, and only 6 % false homologous decisions. As pointed out in experiment 1, this threshold value should be tuned accordingly to the completion rate of database and to the number of tested candidates by the matching algorithm.

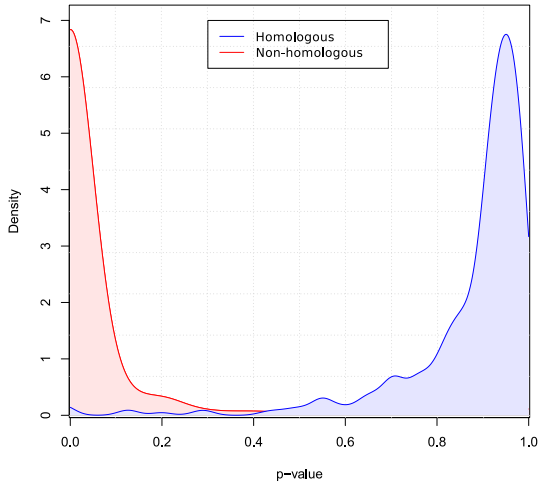


**Figure 19.** Distributions of the p-values estimated by the customized simulations for the homologous (red) and non-homologous (blue) pairs of buildings.

### 6.2. *Discussion*

New insights are acquired thanks to the three experiments allowing us define some recommendation for the practitioner.

The first insight regards the range of the *radial distance* values. Theoretically, the range is comprised between 0 and $\sqrt{2}$, which allows to compute a similarity measure. Our experiments revealed that values of 0.4 already correspond to very different shapes. For a user, without additional information *a priori*, it is advised to consider the range of values of the *radial distance* as evolving between 0 (identical polygons) and 0.5 (very dissimilar polygons).

The second insight deals with the threshold that can be set to separate homologous and non-homologous pair of features. Without additional information, and on the particular case of the OSM - BD TOPO data matching, a threshold equal to 0.07 allows to discriminate the pairs of homologous and non-homologous buildings, with a risk of error estimated at 10% for both false positives and false negatives classes. The first results showed that the distribution of *radial distance* values of non-homologous buildings does not seem to be too sensitive to the radius used to select the candidates for the features to be matched (*i.e.* close candidate buildings located within the search radius area are as similar or dissimilar in terms of *radial distance* as candidates that are far from the building to be matched). These results need to be confirmed with further tests. Moreover, we also extended the criterion using a Bayes rule to accommodate different completeness datasets and different numbers of candidates tested. This rule corresponds to different a priori probabilities of finding a homologous

feature for each feature to be matched.

The third insight is related to the effect of noise. Experiments allowed us to determine a quadratic error summation rule, which lets us handle the case where both databases are significantly noisy. We have seen that, for long-range correlations (e.g. referencing error, camera distortion on aerial images, orthorectification error, etc.), the accepted threshold on the *radial distance* evolves linearly between 0.0 and typically 0.1 to 0.2 depending on the noise level. For lower auto correlation noise (*e.g.* low-cost local topometric survey, automatic detection on Digital Surface Model, building mapping with low cost GPS), the allowed threshold increases rapidly to a plateau value between 0.3 and 0.4. The effect of the noise due to generalization process is more singular. We have observed a fairly marked effect of simplification algorithms (*e.g.* Douglas-Peucker and Vis Valingam), in general at least as important or even more important than the effect of random geometric noise. This may be due to a coordinated and asymmetric effect of the perturbation, which is almost never found in random noise. On the contrary, the effect of squaring is marginal, even in some extreme cases where building shapes are more sensitive to the squaring process.

Fourth, we have shown that *ad hoc* simulations on a given building provide a more discriminating decision threshold (homologous/non-homologous), at the expense of a slight increase in the computation time. In the OSM and BD TOPO data matching case, an error rate of less than 10% is obtained. This approach requires the implementation of an inference of the error model on both databases and a simulation on each individual building to be matched.

Finally, the results of the experiments 1, 2, 3, and 4 allow the user to choose an approach according to the available data and resources. For an OSM and BD TOPO data matching, or with 2 databases having the same types of noise, the abacus of experiment 1 can be used. Otherwise, the abacus of experiment 2 is recommended. If it is known that the two databases are significantly noisy, we recommend completing the abacus from the experiment 2 with the one of the experiment 3. Lastly, for a very fine decision, the method proposed in experiment 4 can be used.

## 7. Conclusion

In this paper, we studied the *radial distance* from different perspectives (*i.e.* properties, computation and robustness) in the context of heterogeneous data matching.

We showed that, despite being named *radial distance* in the literature, it is not a distance from a mathematical point of view. Indeed, we demonstrated it lacks separation, *i.e.* two polygons having different shapes can be separated by a null distance. This implies a contradiction with the interpretation of zero shape distance and in the case of data matching, it can generate incorrect matching links. However, the negative impact is partly offset by two aspects. First, we demonstrated that the separation remains valid for a specific type of polygon (namely star-shaped polygons). This is a relevant result, knowing that the majority of polygons are star-shaped, and as our experiments showed, even non-star-shaped polygons are in fact almost so in terms of proportion of their boundary. Second, the triangle inequality remains

valid and we indicated through an example the importance of this property for data matching.

In terms of computation, we proposed two improvements. The first allows the computation of a radial signature with a fine (potentially infinite) sampling. When defining an operational context, the naive approach of the calculation does not allow to deal with polygons (or signatures) of more than 20 points. With the FFT approach, the problem can be handled with signatures of more than 10 000 points.

Regarding the robustness of the *radial distance*, we conducted a set of theoretical analysis and experiments showing the behavior of the *radial distance* according to different types and amplitudes of noise. Moreover, with the simulation methods, we can easily construct abacuses allowing to decide whether the matching hypothesis should be rejected, with a controlled risk of error, and according to the type and level of noise contaminating the database in which the data matching algorithm is looking for a matching candidate. The last experiment, shown that simulations methods allow to accept or reject more confidently the null hypothesis for a specific pair of features. A refined statistical estimation of the quality of this customized decision algorithm will be the subject or future works.

In addition to the matching context, all our results may be important also for other application fields potentially using the *radial distance*: geographic data quality (*e.g.* when estimating the correctness of a shape knowing a reference feature), multisource data fusion (for reducing errors when fusing different homologous features stemming from different sources), shape classification, generalization and computer vision.

One of the limitations of the study is that the error due to the difference between the nominal terrain between two databases to match is not taken into account. This error is difficult to estimate in the absence of ground truth data, information about the tools used for mapping or lack of specification as it often the case of VGI. Once the error is modeled, one future work is to develop robust methods to make the *radial distance* less sensitive to the noise regarding the nominal terrain. Another potential perspective would be to extend the radial distance to deal with vector-image comparisons.

All experiments were performed on polygons representing buildings. It will be interesting to make similar study on other objects such as parcels, lakes, forests, spatial tesselations, etc. As the radial distance relies on the parametrization of the boundary, it should be a pertinent indicator even for boundaries rougher than typical building edges, provided that level of detail is homogeneous along the boundary. Moreover, the *radial distance* cannot be applied as such for comparing polygons containing holes and more research is needed to make it compliable with this extended type of polygon.

Eventually, managing complex links having $1 : n$ and $m : n$ cardinalities was way beyond the scope of this paper, but may be of paramount importance for leveraging the results found here and the capabilities of the *radial distance* to general polygon data matching algorithms and for all types of situations encountered in practice. One way to manage links with $1 : n$ and $m : n$ cardinalities would be to build new abacuses and derive operative thresholds to determine whether two groups of buildings are homologous (the thresholds computed in this paper may be modified by the fact that

allowing the grouping of buildings may lower computed *radial distances* in general and thus also lower the optimal threshold. A more ambitious way to manage $1 : n$ and $m : n$ cardinalities would be to isolate the effect of a change in the origin of the *radial signature* to identify when two sections of *radial signatures* (that is, restrictions to sub-intervals of $[0, 1]$) correspond to homologous features. Such an analysis could be facilitated by taking into account the typical shapes of buildings, with straight edges and right angles, which constrains the form of the radial signature and could make partial boundary matching more straightforward.

## Notes on contributor(s)

**Yann Méneroux** is a researcher at the French National Mapping Agency. His research interest focuses on GPS trajectories mainly through terrestrial vehicle trajectography with sensor fusion, and usage of collected GPS traces for map construction. His research extends as well to noise analysis and modeling error propagation in applications relying on geographic data.

**Ibrahim Maidaneh Abdi** graduated as a State Engineer in Geoinformation, teaches Geomatics at the University of Djibouti and is currently completing his PhD at the IGN on the "Evaluation of the quality of the OpenStreetMap database with machine learning: Case of the Republic of Djibouti". His main contribution is the proposal of a general framework allowing to infer the extrinsic quality of a building dataset based only on an intrinsic evaluation, in the context of the absence of reference in most countries in Africa.

**Arnaud Le Guilcher** is a researcher in GIS at the LASTIG laboratory. His research interest include data quality, data enrichment, propagation of uncertainties in complex problems involving geographical data, and anonymization, with a specific focus on VGI data.

**Ana-Maria Olteanu-Raimond** is senior researcher in GIS and co-director of LASTIG laboratory. Her research interests include the integration of heterogeneous spatial data, imperfect information fusion, data matching, and collaborative mechanisms and qualification of VGI for joint use with authoritative spatial data. She is also conducting research on citizen science: data collection, platforms, tasks, motivation and sustainability.

All the authors participated to the conceptualization, methodology, validation and writing tasks. A.M.O.R produced the state of the art. A.L.G and Y.M. conducted the formal analysis. Codes and scripts were written by I.M.A and Y.M. Eventually, I.M.A. and A.M.O.R. prepared and anotated the data.

## Data and Codes Availability Statement

The data and code that support the findings of this study are available through the following private link:

`https://doi.org/10.5281/zenodo.7006944`

## Disclosure statement

The authors report there are no competing interests to declare.


## Funding

This work was supported by the French National Mapping Agency: Institut National de l'Information Géographique et Forestière (IGN) and by the University of Djibouti.


## References


Adoram, M. and Lew, M.S., 1999. Irus: image retrieval using shape. *In*: *Proceedings IEEE International Conference on Multimedia Computing and Systems*. IEEE, vol. 2, 597–602.

Al-Bakri, M. and Fairbairn, D., 2012. Assessing similarity matching for possible integration of feature classifications of geospatial data from official and informal sources. *International Journal of Geographical Information Science*, 26 (8), 1437–1456. Available from: https://doi.org/10.1080/13658816.2011.636012.

Angel, S., Parent, J., and CIVCO, D.L., 2010. Ten compactness properties of circles: measuring shape in geography. *The Canadian Geographer / Le Géographe canadien*, 54 (4), 441–461. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1541-0064.2009.00304.x.

Arkin, E.M., *et al.*, 1991. *An efficiently computable metric for comparing polygonal shapes*. CORNELL UNIV ITHACA NY.

Basaraner, M. and Cetinkaya, S., 2017. Performance of shape indices and classification schemes for characterising perceptual shape complexity of building footprints in gis. *International Journal of Geographical Information Science*, 31 (10), 1952–1977. Available from: https://doi.org/10.1080/13658816.2017.1346257.

Buttenfield, B.P., 1991. *Map generalization: Making rules for knowledge representation*.

Chang, C.C., Hwang, S., and Buehrer, D.J., 1991. A shape recognition scheme based on relative distances of feature points from the centroid. *Pattern recognition*, 24 (11), 1053–1063.

Cohen, S.D. and Guibas, L.J., 1997. Shape-based image retrieval using geometric hashing. *In*: *Proceedings of the ARPA Image Understanding Workshop*. ARPA, 669–674.

Costes, B. and Perret, J., 2019. A hidden markov model for matching spatial networks. *Journal of Spatial Information Science*, 2019 (18), 57–89.

Douglas, D.H. and Peucker, T.K., 1973. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: the international journal for geographic information and geovisualization*, 10 (2), 112–122.

Fan, H., *et al.*, 2014. Quality assessment for building footprints data on openstreetmap. *International Journal of Geographical Information Science*, 28 (4), 700–719.

Fonte, C.C., *et al.*, 2015. VGI QUALITY CONTROL. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, II-3/W5, 317–324. Available from: http://www.isprs-ann-photogramm-remote-sens-spatial-inf-sci.net/II-3-W5/317/2015/.

Fonte, C.C., *et al.*, 2017. Using openstreetmap to create land use and land cover maps: Development of an application. *In*: C. Campelo, C. Elízio, M. Bertolotto and P. Corcoran, eds. *Volunteered geographic information and the future of geospatial data*. Hershey, PA: IGI Global, 113–137. Available from: https://doi.org/10.4018/978-1-5225-2446-5.ch007.

Foody, G.M., 2009. Sample size determination for image classification accuracy assessment and comparison. *International Journal of Remote Sensing*, 30 (20), 5273–5291. Available from: https://doi.org/10.1080/01431160903130937.

Fu, Z.L., Shao, S.W., and Tong, C.Y., 2010. Multi-scale area entity shape matching based on tangent space. *Computer Engineering*, 17, 074.

Goodchild, M., 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69, 211–221.

Hangouët, J.F., 2006. Spatial data quality assessment and documentation. *Fundamentals of Spatial Data Quality*, 211–235.

Ivanovic, S., *et al.*, 2019a. A Filtering-Based Approach for Improving Crowdsourced GNSS Traces in a Data Update Context. *ISPRS International Journal of Geo-Information*, 8 (9), 380. Available from: https://www.mdpi.com/2220-9964/8/9/380.

Ivanovic, S.S., *et al.*, 2019b. Potential of Crowdsourced Traces for Detecting Updates in Authoritative Geographic Data. *In*: P. Kyriakidis, D. Hadjimitsis, D. Skarlatos and A. Mansourian, eds. *Geospatial Technologies for Local and Regional Development*. Cham: Springer International Publishing, 205–221. Series Title: Lecture Notes in Geoinformation and Cartography, Available from: http://link.springer.com/10.1007/978-3-030-14745-7_12.

Kim, W.Y. and Kim, Y.S., 2000. A region-based shape descriptor using zernike moments. *Signal processing: Image communication*, 16 (1-2), 95–102.

Liu, C., *et al.*, 2015. A progressive buffering method for road map update using openstreetmap data. *ISPRS International Journal of Geo-Information*, 4 (3), 1246–1264. Available from: https://www.mdpi.com/2220-9964/4/3/1246.

Liu, L., *et al.*, 2021. A data fusion-based framework to integrate multi-source vgi in an authoritative land use database. *International Journal of Digital Earth*, 14 (4), 480–509. Available from: https://doi.org/10.1080/17538947.2020.1842524.

Lokhat, I. and Touya, G., 2016. Enhancing building footprints with squaring operations. *Journal of Spatial Information Science*, 2016 (13), 33–60.

Maidaneh Abdi, I., Le Guilcher, A., and Olteanu-Raimond, A.M., 2020. A regression model of spatial accuracy prediction for openstreetmap buildings. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, 5 (4).

Meng, Q. and Lu, Y., 2014. A fast multi-scale polygon features update approach based on features matching. *In*: *2014 The Third International Conference on Agro-Geoinformatics*. IEEE, 1–5.

Müller, M., 2007. Dynamic time warping. *Information retrieval for music and motion*, 69–84.

Mustière, S. and Devogele, T., 2008. Matching networks with different levels of detail. *GeoInformatica*, 12 (4), 435–453.

Olteanu-Raimond, A.M., *et al.*, 2020. Use of Automated Change Detection and VGI Sources for Identifying and Validating Urban Land Use Change. *Remote Sensing*, 12 (7), 1186. Available from: https://www.mdpi.com/2072-4292/12/7/1186.

Olteanu-Raimond, A.M., Mustière, S., and Ruas, A., 2015. Knowledge formalization for vector data matching using belief theory. *Journal of Spatial Information Science*, (10). Available from: http://josis.org/index.php/josis/article/view/194.

Premaratne, P. and Premaratne, M., 2014. Image matching using moment invariants. *Neurocomputing*, 137, 65–70.

Ripley, B.D., 2009. *Stochastic simulation*. vol. 316. John Wiley & Sons.

Schultz, M., *et al.*, 2017. Open land cover from openstreetmap and remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, 63, 206–213. Available from: https://www.sciencedirect.com/science/article/pii/S0303243417301605.

See, L., *et al.*, 2016. Crowdsourcing, Citizen Science or Volunteered Geographic Information? The Current State of Crowdsourced Geographic Information. *ISPRS International Journal*

*of Geo-Information*, 5 (5), 55. Available from: http://www.mdpi.com/2220-9964/5/5/55.
Senaratne, H., *et al.*, 2017. A review of volunteered geographic information quality assessment methods. *International Journal of Geographical Information Science*, 31 (1), 139–167.
Van Damme, M.D., Olteanu-Raimond, A.M., and Méneroux, Y., 2019. Potential of crowdsourced data for integrating landmarks and routes for rescue in mountain areas. *International Journal of Cartography*, 5 (2-3), 195–213. Available from: https://www.tandfonline.com/doi/full/10.1080/23729333.2019.1615730.
van Winden, K., Biljecki, F., and van der Spek, S., 2016. Automatic update of road attributes by mining gps tracks. *Transactions in GIS*, 20 (5), 664–683. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/tgis.12186.
Vauglin, F. and Bel Hadj Ali, A., 1998. Geometric matching of polygonal surfaces in giss. *In*: *Proc. ASPRS Annual Meeting*.
Visvalingam, M. and Whyatt, J.D., 1993. Line generalisation by repeated elimination of points. *The cartographic journal*, 30 (1), 46–51.
Walter, V. and Fritsch, D., 1999. Matching spatial data sets: a statistical approach. *International Journal of geographical information science*, 13 (5), 445–473.
Xavier, E.M.A., Ariza-López, F.J., and Ureña Cámara, M.A., 2016. A survey of measures and methods for matching geospatial vector datasets. *ACM Comput. Surv.*, 49 (2). Available from: https://doi.org/10.1145/2963147.
Yan, Y., *et al.*, 2020. Volunteered geographic information research in the first decade: a narrative review of selected journal articles in giscience. *International Journal of Geographical Information Science*, 34 (9), 1765–1791. Available from: https://doi.org/10.1080/13658816.2020.1730848.
Zhang, D. and Lu, G., 2004. Review of shape representation and description techniques. *Pattern Recognition*, 37 (1), 1–19. Available from: https://www.sciencedirect.com/science/article/pii/S0031320303002759.
Zielstra, D. and Hochmair, H.H., 2011. Comparative study of pedestrian accessibility to transit stations using free and proprietary network data. *Transportation Research Record*, 2217 (1), 145–152. Available from: https://doi.org/10.3141/2217-18.

## *Appendix A. Proof of Proposition 3.7.*

Proposition 3.7 can be rewritten in the following way:

$$\forall \gamma, \gamma' \in \mathcal{S}, d_r(\gamma, \gamma') = 0 \Rightarrow I(\gamma) = H(I(\gamma'))$$

where $I(\gamma)$ is the interior of the closed curve $\gamma$ and $H$ is a homothety of $\mathbb{R}^2$.

For $s \in [0, L]$, we note $\varphi_\gamma(s)$ the angle between $(C(\gamma), \gamma(0))$ and $(C(\gamma), \gamma(s))$. We choose $\varphi$ such that $\varphi(0) = 0$ and $\varphi$ is continuous on $[0, L]$. As $\gamma$ does not self-intersect, it results that $\varphi(L) \in \{-2\pi, 0, 2\pi\}$. $I(\gamma)$ being star-shaped around $C(\gamma)$ implies that for all $\theta \in [0, 2\pi]$, $\varphi_\gamma^{-1}(\theta)$ is an interval of $[0, L[$. As a consequence, $\varphi(s)$ is either nondecreasing or nonincreasing. As we also suppose that $\gamma$ turns anticlockwise, it implies that $\varphi$ is a nondecreasing function of $s$ and $\varphi(L) = 2\pi$.
Let us also consider the function:

$$\begin{array}{rccc} \rho: & [0, L] & \to & \mathbb{R} \\ & s & \mapsto & \|\gamma(s) - C(\gamma)\| \end{array}$$

If $\gamma$ is a polygon and $V(\gamma)$ the set of its vertices, then $\varphi$ and $\rho$ are piecewise-

differentiable on $[0, L]$; the only points where they not differentiable are in the set $\gamma^{-1}(V(\gamma))$, and $\varphi$ and $\rho$ are semi-differentiable at these points.

We consider $s_0 \in [0, L[$ and study the right derivatives of $\varphi$ and $\rho$ at point $s_0$ (they are written $\varphi'_r(s_0)$ and $\rho'_r(s_0)$. We also consider $s_1 > s_0$ such that

$$]s_0, s_1[ \ \cap \ \gamma^{-1}(V(\gamma)) = \varnothing.$$

Then, $\gamma([s_0, s_1])$ is the segment $[\gamma(s_0), \gamma(s_1)]$, and for all $s \in [s_0, s_1]$, it results that:

$$(s - s_0)^2 = (\rho - \rho_0)^2 + \left(2\rho \sin\left(\frac{\varphi - \varphi_0}{2}\right)\right)^2$$

thus, it results that:

$$\sin\left(\frac{\varphi - \varphi_0}{2}\right)^2 = \frac{1}{4\rho^2}\left((s - s_0)^2 - (\rho - \rho_0)^2\right) = (s - s_0)^2 \frac{1}{4\rho^2}\left(1 - \left(\frac{\rho - \rho_0}{s - s_0}\right)^2\right)$$

We consider the limit of this equality when $s \to s_0$:

$$\begin{array}{ll} \lim_{s \to s_0} \dfrac{\sin\left(\dfrac{\varphi - \varphi_0}{2}\right)^2}{(s - s_0)^2} & = \left(\dfrac{1}{2}\varphi'_r(s_0)\right)^2 \\ \lim_{s \to s_0} \dfrac{1}{4\rho^2}\left(1 - \left(\dfrac{\rho - \rho_0}{s - s_0}\right)^2\right) & = \dfrac{1}{4\rho^2}(1 - \rho'_r(s_0)^2) \end{array}$$

and the following expression is inferred:

$$\varphi'_r(s_0)^2 = \frac{1 - \rho'_r(s_0)^2}{\rho(s_0)^2}$$

As $\varphi$ and $\rho$ are continuous and differentiable almost everywhere, we have

$$\varphi(s) = \int_0^s \frac{\sqrt{1 - \rho'(\sigma)^2}}{\rho(\sigma)} d\sigma$$

As a consequence, the function $s \to \rho(s)$ determines the values of $\varphi(s)$ for all $s \in [0, L]$. So, if $\gamma_1$ and $\gamma_2$ are such that $\rho_1 = \rho_2$, it is approved that $\varphi_1 = \varphi_2$.

Now, let us consider $\gamma_1$ and $\gamma_2$ such that $r_{\gamma_1} = r_{\gamma_2}$. For all $t \in [0, 1]$, it results that:

$$\frac{\rho_1(tL_1)}{R_1} = \frac{\rho_2(tL_2)}{R_2}$$

.

That implies:

$$\begin{aligned}\varphi_1'\left(tL_1\right)^2 &= \frac{1-\rho_1'\left(tL_1\right)^2}{\rho_1\left(tL_1\right)^2} \\ &= \frac{1-\left(\frac{L_2R_1}{L_1R_2}\right)^2\rho_2'\left(tL_2\right)^2}{\left(\frac{R_1}{R_2}\right)^2\rho_2\left(tL_2\right)^2} \\ &= \left(\frac{L_2}{L_1}\right)^2\frac{\left(\frac{L_1}{L_2}\right)^2-\left(\frac{R_1}{R_2}\right)^2\rho_2'\left(tL_2\right)^2}{\left(\frac{R_1}{R_2}\right)^2\rho_2\left(tL_2\right)^2}\end{aligned}$$

Now, we suppose that $\frac{L_1}{L_2} > \frac{R_1}{R_2}$. This implies $\varphi_1'\left(tL_1\right)^2 > \left(\frac{L_2}{L_1}\right)^2\varphi_2'\left(tL_2\right)^2$. As $\varphi_1$ and $\varphi_2$ are nondecreasing, this means that $L_1\varphi_1'\left(tL_1\right) > L_2\varphi_2'\left(tL_2\right)$. Then, it results that:

$$2\pi = \varphi_1\left(L_1\right) = \int_0^{L_1}\varphi_1'(s)ds = L_1\int_0^1\varphi'\left(tL_1\right)dt < L_2\int_0^1\varphi'\left(tL_2\right)dt = 2\pi,$$

which brings a contradiction.

Similarly, if $\frac{L_1}{L_2} < \frac{R_1}{R_2}$, then $L_1\varphi_1'\left(tL_1\right) < L_2\varphi_2'\left(tL_2\right)$, which is also a contradiction.

Thus, considering the equality $\frac{L_1}{L_2} = \frac{R_1}{R_2}$, and for all $t \in [0,1]$, $L_1\varphi_1'\left(tL_1\right) = L_2\varphi_2'\left(tL_2\right)$, which implies

$$\forall s \in [0, L_2], \varphi_2(s) = \varphi_1\left(\frac{L_2}{L_1}s\right).$$

The equality $\frac{L_1}{L_2} = \frac{R_1}{R_2}$ also gives

$$\forall s \in [0, L_2], \rho_2(s) = \frac{L_2}{L_1}\rho_1\left(\frac{L_1}{L_2}s\right).$$

This means that, if $C(\gamma_1) = C(\gamma_2) = C$ and $\gamma_2(0) - C = \frac{L_2}{L_1}(\gamma_1(0) - C)$, then

$$\forall s \in [0, L_2], \gamma_2(s) - C = \frac{L_2}{L_1}\left(\gamma_1\left(\frac{L_1}{L_2}s\right) - C\right).$$

and

$$I(\gamma_2) = D_{\left(C, \frac{L_2}{L_1}\right)}\left(I(\gamma_1)\right)$$

where $D_{\left(C,\frac{L_2}{L_1}\right)}$ is the dilation of center $C$ with a factor $\dfrac{L_2}{L_1}$ (or a contraction if the factor is lower than 1).

In general, when $C(\gamma_1) \neq C(\gamma_2)$, then

$$\forall s \in [0, L_2]\,, \begin{pmatrix}\gamma_2^x\\ \gamma_2^y\end{pmatrix}(s) = \begin{pmatrix}C(\gamma_2)^x\\ C(\gamma_2)^y\end{pmatrix} + \frac{L_2}{L_1}\begin{pmatrix}\cos\Theta & -\sin\Theta\\ \sin\Theta & \cos\Theta\end{pmatrix}\begin{pmatrix}\gamma_1^x - C(\gamma_1)^x\\ \gamma_2^y - C(\gamma_1)^y\end{pmatrix}\begin{pmatrix}\frac{L_1}{L_2}s\end{pmatrix}$$

with $\Theta$ the angle between vectors $\overrightarrow{C(\gamma_1)\gamma_1(0)}$ and $\overrightarrow{C(\gamma_2)\gamma_2(0)}$. Then, it results that:

$$I(\gamma_2) = D_{\left(C(\gamma_2),\frac{L_2}{L_1}\right)} \circ R_{(C(\gamma_2),\Theta)} \circ T_{C(\gamma_2)-C(\gamma(1)}\left(I(\gamma_1)\right),$$

with $R_{(C(\gamma_2),\Theta)}$ the rotation of an angle $\Theta$ around the center $C(\gamma)$, and $T_{C(\gamma_2)-C(\gamma(1)}$ being the translation of vector $\overrightarrow{C(\gamma_1)C(\gamma_2)}$.

Finally, we consider $\gamma_1$ and $\gamma_2$ such that $d_r(\gamma_1, \gamma_2) = 0$. We notice that as $r_{\gamma_1}$ and $r_{\gamma_2}$ are continuous, $d_r(\gamma_1, \gamma_2) = 0$ means that there exists $\tau \in [0, 1[$ such that $r_{\gamma_2} = r_{\gamma_1}(\cdot + \tau)$. the function $r_{\gamma_1}(\cdot + \tau)$ is the radial signature of the closed curve:

$$\begin{array}{rccl}\widetilde{\gamma}_1: & [0, L_1] & \to & \mathbb{R}^2\\ & s & \mapsto & \begin{cases}\gamma(s + \tau L_1) & \textbf{if } s < L_1(1-\tau)\\ \gamma(s + (\tau-1)L_1) & \textbf{if } s \geqslant L_1(1-\tau)\end{cases}.\end{array}$$

The curves $\gamma_1$ and $\widetilde{\gamma}_1$ have the same image, so $I(\gamma_1) = I(\widetilde{\gamma}_1)$. As $r_{\gamma_2} = r_{\gamma_1}(\cdot + \tau)$, $I(\gamma_2) = H(I(\widetilde{\gamma}_1)) = H(I(\gamma_1))$, where $H$ is an homothety. ■

### *Appendix B. Proof of Proposition 3.8.*

We exhibit a sequence of closed curves $(\gamma_n)_{n\in\mathbb{N}}$, whose distance to a unit circle $\gamma$, $d(\gamma_n, \gamma)$, converges to 1 as $n$ goes to infinity. Referring to Figure 1, let us define a sequence $(\beta_n)_{n\in\mathbb{N}}$ of *star-shaped* closed curves centered on the origin and composed of a number of radial branches which is an integer multiple[2] of $4n^2$. For example, $\beta_1$ contains 4 branches, $\beta_2$ has 16 branches, $\beta_3$ has 36 branches, and so on. Second, let us define another sequence of *selector* closed curves $(\alpha_n)_{n\in\mathbb{N}}$, composed of a *damped* circle, selecting only the north and south branches of its corresponding $\beta_n$, all other remaining branches being damped by an homothetic factor $\mathcal{O}(n^{-1})$. Eventually, we define the sequence $(\gamma_n)_{n\in\mathbb{N}}$ as a pointwise product (in polar space) of $\alpha$ and $\beta$ sequences (bottom row in Figure 1). Because the length of $\gamma_n$ goes to infinity (as a product of $\mathcal{O}(n^{-1}) \times \mathcal{O}(n^2)$), the abscissa portion of $r_{\gamma_n}$ not decreasing towards 0 is smaller as $n$ grows, and eventually tends to 0. Geometrically, for a large value of $n$, $\gamma_n$ is nearly a straight line (Figure 7). In the radial signature space, it is a double *half-dirac* function $r_{\gamma_n} : t \mapsto \frac{1}{2}(\delta(t) + \delta(t - \frac{1}{2}))$, for which it is not difficult to show that $||r_{\gamma_n} - r_\gamma|| = \sqrt{2}$ and then $d_r(\gamma_n, \gamma) = \sqrt{2}$.

[2]Always being an integer multiple of 4 ensures that all closed curves have perfectly symmetrical north and south branches, and then that their center of mass is invariably located at the origin even after damping selection by $\alpha_n$.

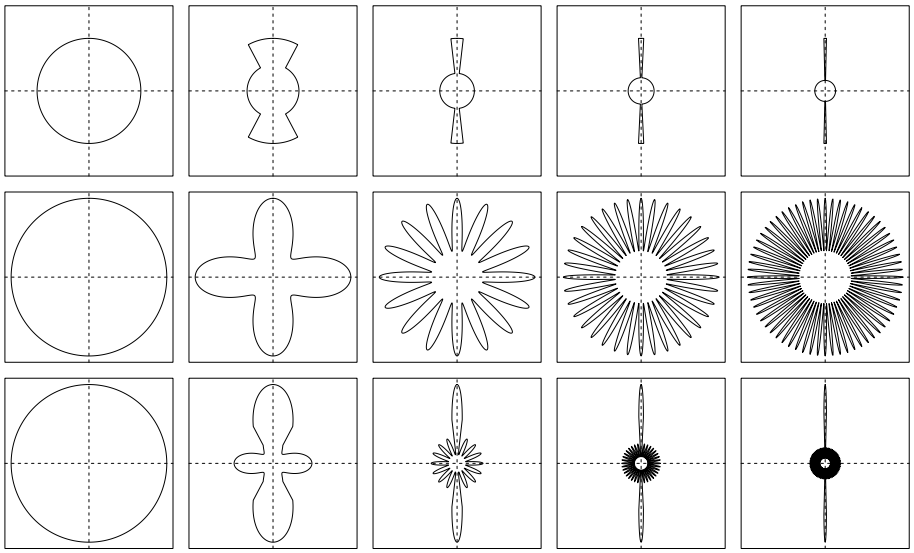

**Figure 1.** First five closed curves of the sequences $(\alpha_n)_{n\in\mathbb{N}}$ (top), $(\beta_n)_{n\in\mathbb{N}}$ (middle) and $(\gamma_n)_{n\in\mathbb{N}}$, (bottom) defined as the element-wise product $\gamma_n = \alpha_n \times \beta_n$. It is not difficult to demonstrate that the length of $\gamma_n$ grows infinitely.

Then, it just remains to show that no pair of radial signatures has distance $\sqrt{2}$. First, let us note that any radial signature computed from a closed curve of length $L$ is $L$-lipschitz continuous. Suppose that $\alpha$ and $\beta$ are two closed curves such that $d_r(\alpha, \beta) = \sqrt{2}$. Then, it means that $\langle r_\alpha(\cdot)|r_\beta(\cdot + \tau)\rangle = 0$ for a particular shift $\tau$, which is possible only if $r_\alpha$ (as well as $r_\beta$) is non-zero on a at most a countable set, which is impossible for any finite-length curve, since $r_\alpha$, of length $L_\alpha$ is $L_\alpha$-lipschitz continuous. Therefore, there is no couple $(\alpha, \beta) \in \Gamma^2$ such that the upper bound $\sqrt{2}$ is reached by $d_r(\alpha, \beta)$ which concludes the proof. ■

The constructive process used in this proof is interesting, since it involves only star-shaped polygons, hence it reveals that any unit-norm $\frac{1}{2}$-periodic piecewise constant function can be approximated by a sequence of radial signatures of curves in $\Gamma^*$. Additionnaly, any unit-norm continuous function of $L^2([0, 1])$ being the simple limit of a sequence of unit-norm piecewise constant functions, the image through the radial signature application of $\Gamma^*$ is a dense subset of $L^2([0, 1])$. Pushing things a little further, we may demonstrate that these functions can actuallay be exactly represented by the radial signature of a star-shaped polygon, which means that $\mathrm{Im}(\Gamma^*) = L^2([0, 1])$, and then the radial signature cannot be an injection on any set greater that $\Gamma^*$. As a consequence, for any non-star-shaped polygon $\gamma \in \Gamma$, it is always possible to find a star-shaped polygon $\gamma^* \in \Gamma^*$, such that $d_r(\gamma, \gamma^*) = 0$, implying that the *radial distance* separation property can be rejected at the outset as it cannot hold outside of $\Gamma^*$. This is an alternate point of view motivating the Proposition 3.7. In some aspects, the set $\Gamma^*$ of star-shaped polygons is a maximal subset of $\Gamma$ (in terms of inclusion) where the *radial distance* is still a mathematical distance.

It is interesting now to discuss how this upper bound $\sqrt{2}$ is decreasing towards 0 as we constrain the radial signatures of polygons to be bounded as it is the case in practice. For $M \geqslant 1$, we consider the subset $\Gamma^M \subset \Gamma$ of polygons $\gamma$ such that $\|r_\gamma\|_\infty \leqslant M$.

Using the shift[3] to reduce the $L^2$-norm between translated signatures provides the following upper bound on the *radial distance* values:

$$\forall\ \alpha, \beta \in \Gamma^M \qquad \begin{cases} d_r(\alpha, \beta) \leqslant \sqrt{2 - \frac{2}{M}} & \text{if } M \leqslant 2 \\ \\ d_r(\alpha, \beta) \leqslant \sqrt{2 - \frac{1}{M^2}} & \text{if } M \geqslant 2 \end{cases}$$

When $M \leqslant 2$, the optimal strategy to minimize the inner product $\langle f \mid g \rangle$ is to take $f = 1$ and have $g$ be a limit of radial signatures, $g$ taking only the values 0 and $M$. When $M \geq 2$, the optimal strategy is to have $f$ and $g$ both be limits of signatures and bivalued, taking only values 0 and $M$, with supports that have the same intersection span for any value of the shift.

This result is important since in practical situations, $M$ does not exceed the value of 2, hence the *radial distance* application may be considered as bounded by 1 (and even then upper bounds are only reached with elaborate radial signatures).

[3]The upper bound $\sqrt{2}$ is approached with signatures that have arbitrary large values and, thus arbitrary small support, which enables arbitrary small value of the inner product $\langle \cdot \mid \cdot \rangle$ even with the optimal shift. When we constrain the radial signatures to be bounded by $M$, their support have a minimum span and an optimal shift can make them overlap and necessarily lowers the upper bound.