PNAS
www.pnas.org

Supplementary Information for

Segregation through the multiscalar lens: focal distances and distortion coefficients

Madalina Olteanu, Julien Randon-Furling and William Clark

Madalina Olteanu
E-mail: madalina.olteanu@univ-paris1.fr

This PDF file includes:

Supplementary text
Figs. S1 to S14

www.pnas.org/cgi/doi/10.1073/pnas.1900192116

Supporting Information Text

In order to bring more insights on the various notions introduced in the manuscript – KL-divergence trajectories, focal distances and focal-distances trajectories, distortion coefficients, normalization procedure –, we illustrate them on various synthetic data. This allows us to detail and clarify all the definitions above, as well as the intrinsic properties of all these notions.
In all cases, we used a square regular grid, in which each unit represents one individual. The size of the grid is meaningless, since one may always normalize with respect to it.
We shall start with various examples for a two-group case, and then extend our analysis to a four-group case.

Experimental results on two-group configurations

Two groups, A and B, with proportions $p_0 < 0.5$ and $1 - p_0$ are considered. More particularly, in the examples below, $p_0 = 0.25$.

Extreme case for a two-group distribution. The first example we present here is the extreme-segregation case for two-group distributions: the two populations are completely separated, and the smallest one is isolated in one "ghetto", here in upper right corner of Fig. S1 (left). This configuration contains the maximally segregated unit – the green point in the extreme upper right corner –, which is generating the KL-divergence trajectory with the largest possible distortion coefficient. The KL-divergence trajectory of their unit, as defined in Eq. 2, is represented in Fig. S1 (right): the individual starting his trajectory from this location will first see all the individuals from the same group as theirs, and only afterwards meet those from the other group. His KL-trajectory will start at the maximum possible value of the KL divergence, $-\log p_0$, remain constant while they keep meeting individuals from the same group, and then slowly converge towards 0.
When one computes the set of KL trajectories for all the units in the grid, it sets the representation in Fig. S2 (left). All trajectories eventually converge to 0, and are dominated by that of the maximally segregated unit. The next step of the analysis consists in analyzing the convergence of these trajectories. By varying a convergence threshold δ between 0 and the maximum value of the KL trajectories, one may build trajectories of focal distances, as defined in Eq. 3, which translate the convergence "instant" – in terms of aggregated population –, for each value of δ. The corresponding focal-distances trajectories for our first example are plotted in Fig. S2 (right), where the maximally segregated unit trajectory is enhanced again. One may easily remark that for the extreme unit, the KL and the focal distance trajectories are identical, modulo an axis switch. The area under these two curves is the maximum among all possible curves and configurations, and will serve as normalizing constant for all distortion coefficients in the subsequent. An analytical expression of the normalizing constant, depending only on p_0 and N, the population size, can be easily derived.
When one computes the areas under the focal-distances trajectories and normalizes them as explained above, one obtains the map of the normalized distortion coefficients, as illustrated in Fig. S3. As one may expect, the upper right corner is the most distorted, with normalized coefficients close to the maximum value.
Since the configuration in Fig. S1 is an extreme segregation case, one easily sees the distorted area in the upper right corner, but the information on the rest of the map is not visible. This is due to the fact that generally distortion coefficients span over several orders of magnitude, and their distribution is heavily tailed (see Fig. S6, Setting 1 corresponds to the example discussed here). For this reason, we complement the representation with a logged map, which enhances, on the contrary, the less distorted area, having the smallest distortion coefficients and hence the fastest convergence towards the city distribution, as shown in Fig. S4.

Two sets of simulated scenarios for two-group distributions. We use the extreme previous configuration as a starting point for simulating two sets of scenarios. The first set of scenarios, illustrated in Fig. S5 and Fig. S6, contains a clustered structure in each of its configurations. This structure, and, more importantly, its intensity in terms of distortion, are highlighted by the normalized distortion coefficients maps, both in linear and logged scales. Furthermore, in the density representation, we add a supplementary information consisting of the average normalized distortion coefficient and its 95% confidence interval in the case of a null model, obtained by random permutations and corresponding to a perfectly random spatial distribution of the units. This allows to perform empirical tests on the random or clustered structure of the data. According to the results in Fig. S6, more than 95% of the normalized distortion coefficients are larger than the right limit of the confidence interval, so the null hypothesis of a random spatial distribution is systematically rejected for the first scenario. Clustered structure is found in each case, but the distributions of the normalized distortion coefficients and their levels are quite different: the first configuration is the extremely segregated one, it has an important dispersion and is bi-modal; the fourth setting appears as the least segregated and the most homogeneous among the five.
In the second scenario illustrated in Fig. S7 and Fig. S8, the synthetic configurations start with the same extremely segregated case and evolve towards a very regular structure. This translates into KL-trajectories and focal-distance trajectories which converge very fast towards 0. Furthermore, on the last two configurations, the maps of the distortion coefficients do not show any atypical areas, and the differences are only visible on the logged representation.
When plotting the estimated densities of the normalized distortion coefficients, one can easily see that we have here two extreme cases. On the one hand, the very flat, highly dispersed distribution of the first, extremely segregated configuration; on the other hand, the peaked, almost dispersion-free distribution of the last, perfectly regular configuration. When comparing with the null model, the random distribution hypothesis is rejected for the first two configurations, but not for the last three. In the third setting, 24% of the distortion coefficients lie within the margins of the confidence interval, whereas 76% lie at its right; in the fourth setting, 12% lie at its right and 81% within its limits; eventually, in the last case, 56% fall within the confidence

interval and 44% at its left. These numbers suggest an investigation of whether the right-hand side of the confidence interval for the null model corresponds to segregated structures of the city, whereas the left-hand side corresponds to very fine regular structures. The empirical results on simulations and real data appear to support this hypothesis but, to fully answer it, we should first be able to mathematically characterize the null model, which is not completely straightforward yet, as mentioned in the Conclusion of the manuscript.

Experimental results on multi-group configurations

After having illustrated the main concepts on a two-group distribution, let us now generalize to the multi-group case. We shall proceed similarly as above, by introducing the extreme segregation configuration, which contains the maximally segregated unit used as a reference in the normalization procedure. In order to connect the Supplementary Material with the main manuscript, we shall use here a four-group distribution, with proportions relatively close to those of the real data on ethnic mixing in Los Angeles: $p_1 = 0.09$, $p_2 = 0.16$, $p_3 = 0.24$ and $p_4 = 0.51$.

Extreme case for a four-group distribution. As mentioned in the manuscript, the most segregated individual would live in a configuration where the four groups are ordered by size, and where that person would first meet all the individuals of their own group, then all those of the second most unfrequent group, and so on, until having seen the entire population of the city. The KL-divergence trajectory of such an individual would be that illustrated in Fig. S9 (right). If one uses the nearest-neighbor rule for the aggregation procedure and for computing the KL-divergence trajectory, than the configuration which would contain this particular individual would be that in Fig. S9 (left).
Similarly to the two-group case, one may then compute the set of KL-divergence trajectories, and the set of focal-distances trajectories, as in Fig. S10. All KL trajectories eventually converge to 0, and are dominated by that of the maximally segregated unit. Most of the comments and remarks we made for the two-group distribution hold also in this case. The area under the focal-distance curve of the maximally segregated unit will be used hereafter as normalizing constant. An explicit analytical form depending on N, p_1, p_2, p_3 and p_4 may be derived also.
The corresponding normalized distortion coefficients are mapped in Fig. S11, while their logged representation is provided in Fig. S12. On both maps, one may easily identify the most distorted and the less distorted, *clear viewpoints*, of this configuration.

A set of simulated scenarios for multi-group distributions. We use the previous extreme configuration as starting point for several other simulated configurations, while preserving the same groups proportions. The structure and in particular the intensity of segregation at any point in the grid are well highlighted by the normalized distortion coefficients maps in Fig. S13, both in the linear and logged scale. The logged scale appears as better suited to identify the regions at which the perception of the city is less distorted.
When comparing the distributions of the normalized distortion coefficients as in Fig. S14, one may notice some similarities. All configurations have a clustered structure, hence distortion coefficients have rather large values, and their distributions are generally bi-modal, heavy-tailed, and with an important dispersion. Furthermore, when performing random permutations of the units and building the average and the 95% confidence interval of a null model, one may easily see that the null hypothesis is systematically rejected by the empirical test. Our method is thus able to test segregation against random spatial distribution, but, more importantly, is able to identify the sensitive regions where segregation occurs, and the intensity of it.

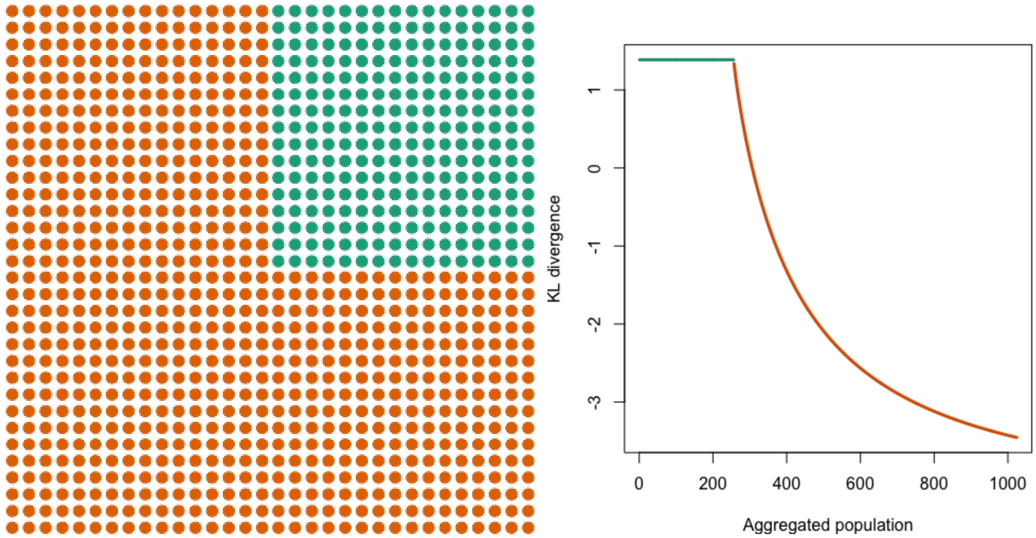

Fig. S1. Two-group simulated scenario with extreme segregation. Left: city configuration, $p_0 = 0.25$. Right: the KL-trajectory of the most segregated unit (located at the upper right corner in the left image).

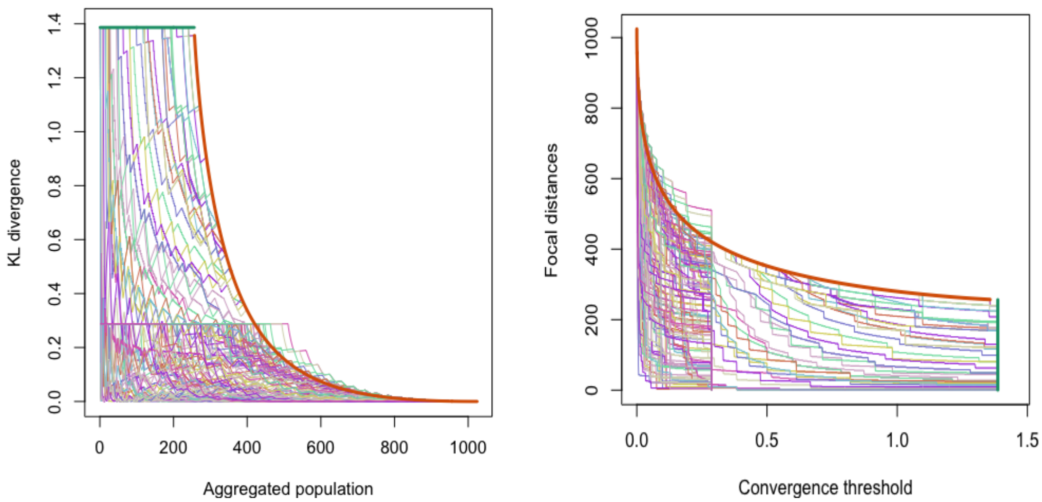

Fig. S2. Two-group simulated scenario with extreme segregation. Left: KL-trajectories (the maximally segregated unit trajectory is enhanced). Right: Focal-distances trajectories (the maximally segregated unit trajectory is enhanced). Distortion coefficients are computed as the areas under the focal-distances trajectories.

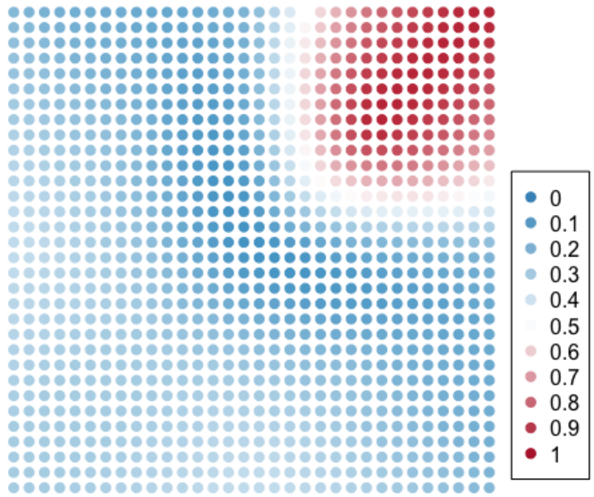

Fig. S3. Two-group simulated scenario with extreme segregation: normalized distortion coefficients.

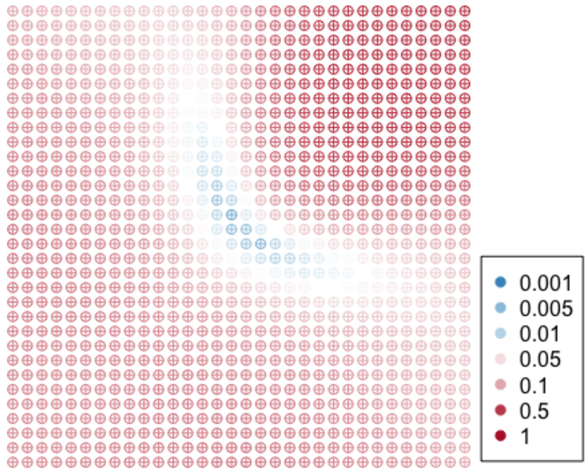

Fig. S4. Two-group simulated scenario with extreme segregation: logged representation of the normalized distortion coefficients

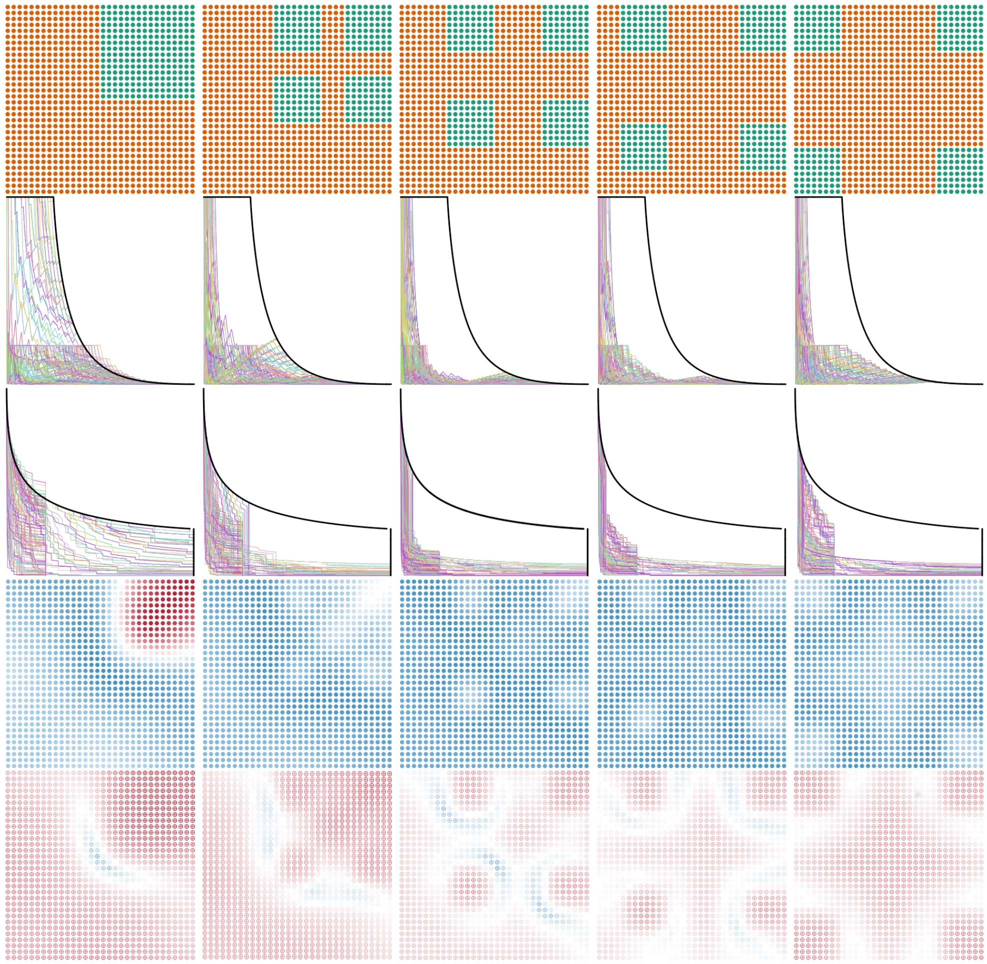

Fig. S5. Scenario 1: various examples of a two-group distribution. Each column represents a different configuration, while overall group proportions are unchanged, $p_0 = 0.25$. First row: simulated spatial configurations. Second row: KL trajectories (the black solid line corresponds to the maximally segregated unit in the extreme segregation scenario). Third row: focal-distances trajectories (the black solid line corresponds to the maximally segregated unit in the extreme segregation scenario, used as reference for the normalization). Fourth row: normalized distortion coefficients (the color scale is that of Fig. S3). Fifth row: logged representation of the normalized distortion coefficients (the color scale is that of Fig. S4).

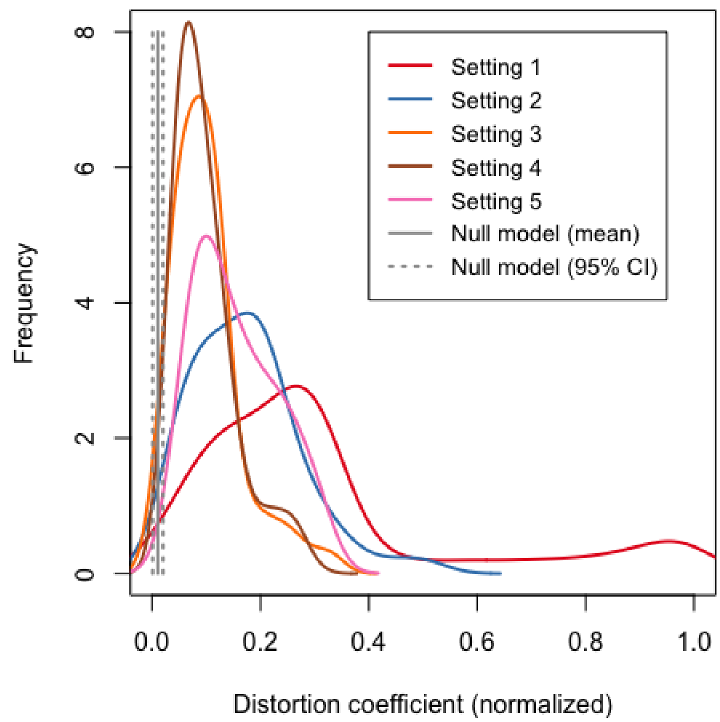

Fig. S6. Scenario 1: estimated densities of the normalized distortion coefficients. The dark grey solid vertical line corresponds to the average value of distortion coefficients in a *null model* obtained by random (spatial) permutations. Dashed lines indicate a 95% interval around the average for the frequency distribution of distortion coefficients in the null model.

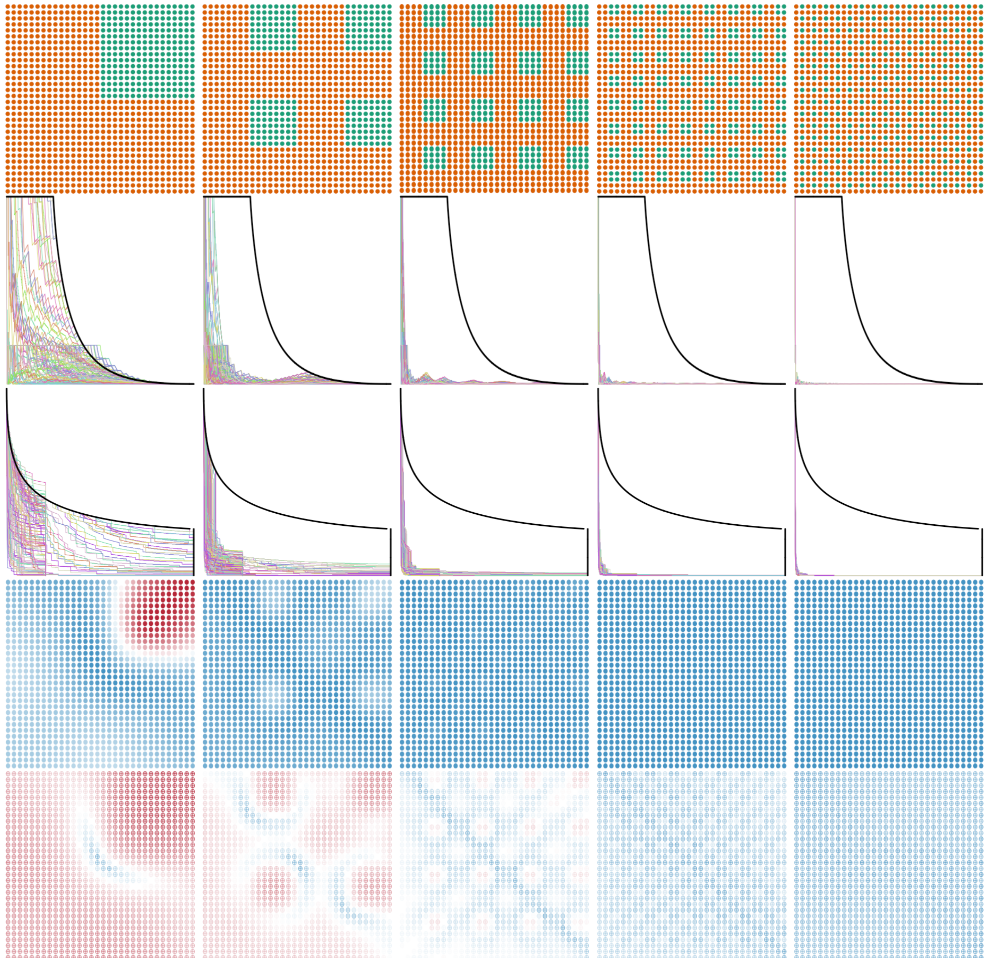

Fig. S7. Scenario 2: various examples of a two-group distribution. Each column represents a different configuration, while overall group proportions are unchanged, $p_0 = 0.25$. First row: simulated spatial configurations. Second row: KL trajectories (the black solid line corresponds to the maximally segregated unit in the extreme segregation scenario). Third row: focal-distances trajectories (the black solid line corresponds to the maximally segregated unit in the extreme segregation scenario, used as reference for the normalization). Fourth row: normalized distortion coefficients (the color scale is that of Fig. S3). Fifth row: logged representation of the normalized distortion coefficients (the color scale is that of Fig. S4).

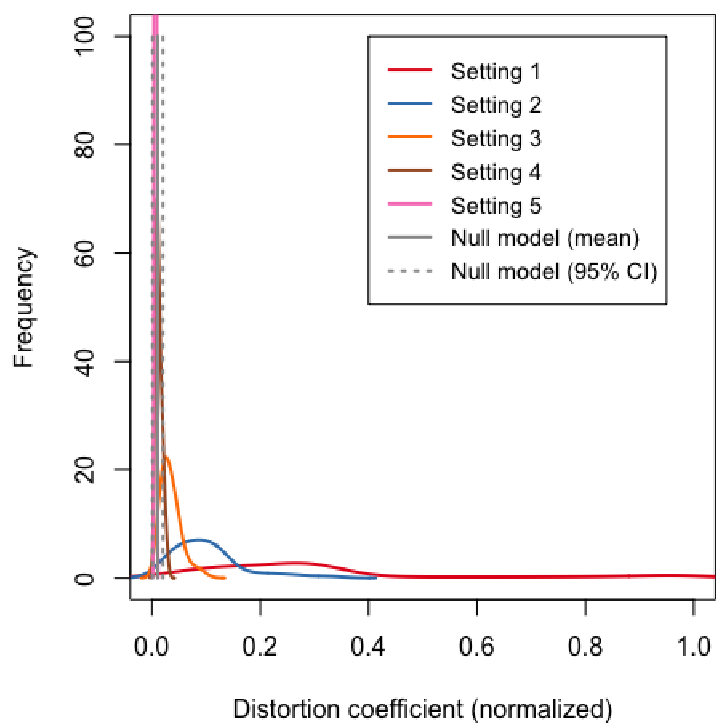

Fig. S8. Scenario 2: estimated densities of the normalized distortion coefficients. The dark grey solid vertical line corresponds to the average value of distortion coefficients in a *null model* obtained by random (spatial) permutations. Dashed lines indicate a 95% interval around the average for the frequency distribution of distortion coefficients in the null model.

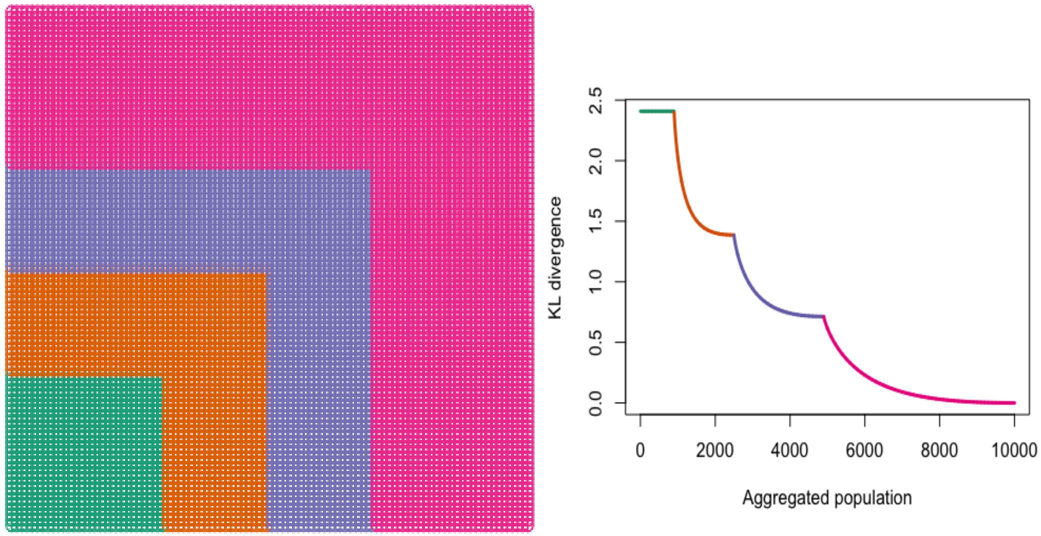

Fig. S9. Four-group simulated scenario with extreme segregation. Left: city configuration, $p_1 = 0.09$ (green), $p_2 = 0.16$ (orange), $p_3 = 0.24$ (purple), $p_4 = 0.51$ (fuchsia). Right: the KL-trajectory of the most segregated unit (located at the lower left corner in the left image).

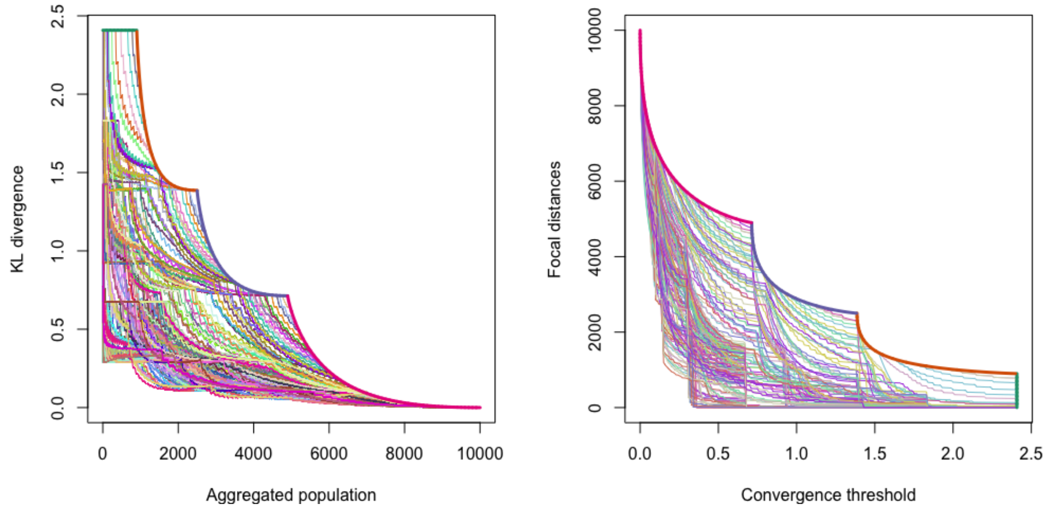

Fig. S10. Four-group simulated scenario with extreme segregation. Left: KL-trajectories (the maximally segregated unit trajectory is enhanced). Right: Focal-distances trajectories (the maximally segregated unit trajectory is enhanced). Distortion coefficients are computed as the areas under the focal-distances trajectories.

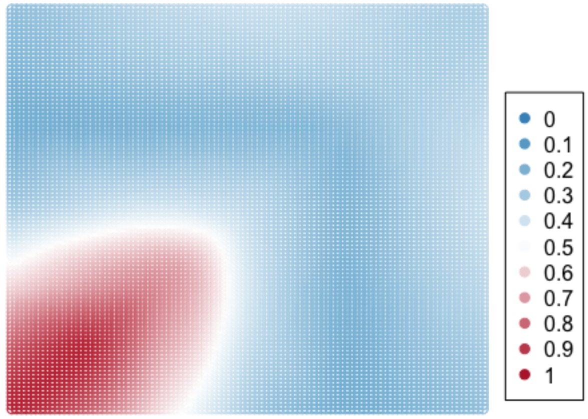

Fig. S11. Four-group simulated scenario with extreme segregation: normalized distortion coefficients.

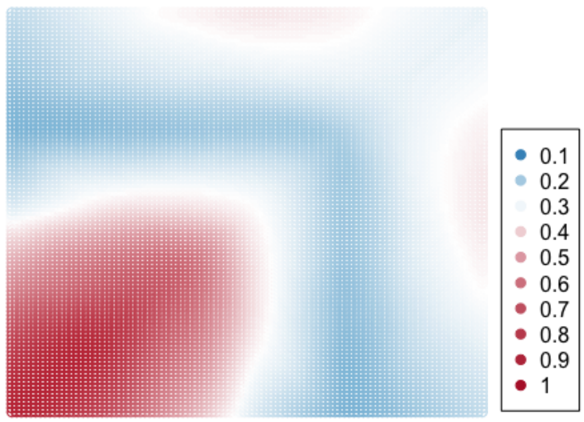

Fig. S12. Four-group simulated scenario with extreme segregation: logged representation of the normalized distortion coefficients

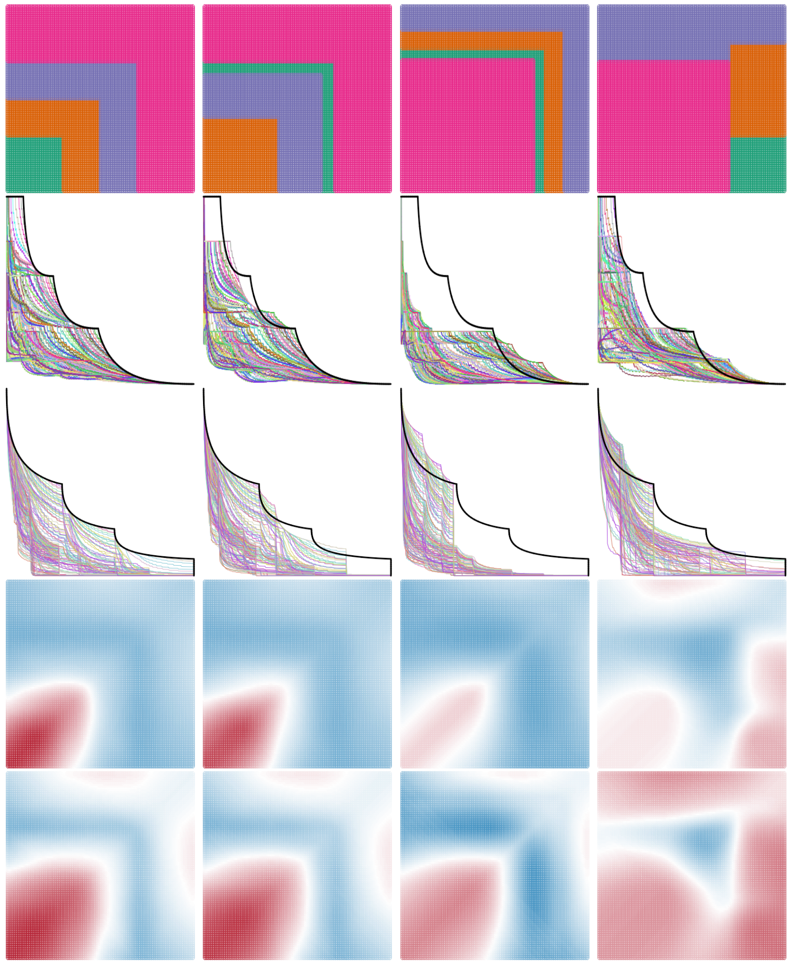

Fig. S13. Scenario 3: various examples of a four-group distribution. Each column represents a different configuration, while overall group proportions are unchanged, $p_1 = 0.09$ (green), $p_2 = 0.16$ (orange), $p_3 = 0.24$ (purple), $p_4 = 0.51$ (fuchsia). First row: simulated spatial configurations. Second row: KL trajectories (the black solid line corresponds to the maximally segregated unit in the extreme segregation scenario). Third row: focal-distances trajectories (the black solid line corresponds to the maximally segregated unit in the extreme segregation scenario, used as reference for the normalization). Fourth row: normalized distortion coefficients (the color scale is that of Fig. S11). Fifth row: logged representation of the normalized distortion coefficients (the color scale is that of Fig. S12).

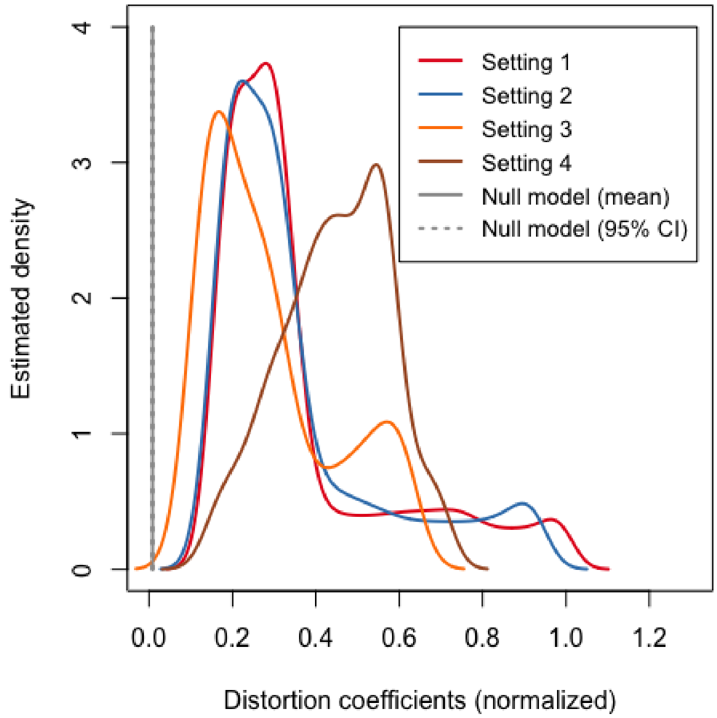

Fig. S14. Scenario 3: estimated densities of the normalized distortion coefficients. The dark grey solid vertical line corresponds to the average value of distortion coefficients in a *null model* obtained by random (spatial) permutations. Dashed lines indicate a 95% interval around the average for the frequency distribution of distortion coefficients in the null model.