The transformed optimal transportation problem: sensitivity and segregation of the children-to-school constrained assignment in Lausanne

Théophile Emmanouilidis, Guillaume Guex, François Bavaud

University of Lausanne, Switzerland

*Corresponding author: Theophile.Emmanouilidis@unil.ch

Abstract: We present a fuzzy approach aiming at assigning origins (children home locations) to destinations (schools) while considering capacity and segregation constraints. Solving the classic instance of the optimal transportation problem generally leads to many equivalent solutions, and can be computationally demanding for large networks. Introducing an additional origin-destination mutual information term, measuring the assignment cohesiveness and penalizing hard assignments, allows faster computing of an unique solution. The algorithm is built on two freely adjustable parameters : the temperature T controlling the assignment cohesiveness, and the mixing parameter c, favoring social diversity at the destinations. The approach yields relevant results in school organisation and is currently used by the public administration of Lausanne to plan children-to-school assignment as well as to evaluate and simulate the development of the schooling infrastructures.

Key-words: optimal transport, regularization, fuzzy clustering, fuzzy logic, social mixing modelling, location-allocation uncertainty, spatial decision making, Lagrangian multipliers

I INTRODUCTION

May it be from an urbanistic, a social or from a governance point of view, the evolution of cities is a major challenge of our contemporary societies. By giving the opportunity to analyze spatial and social configurations or attempting to simulate future ones, GIS cannot be overlooked in urban planning and management. In five years, the population of the city of Lausanne has grown from $134'700$ to $140'570$ inhabitants while the numbers in public schools have increased from $12'200$ to $13'500$ schoolchildren. Demographic rise and dynamics constitute spatially heterogenous processes that have a direct impact on school organization and planning. They require the adaptation of the schooling infrastructures and may generate uncertainty in schoolchildren location-allocation. Nowadays most cities have divided their territory into multiple school districts to assign children to school. But sooner or later such districts become inaccurate as they define fixed boundaries in an evolving urban landscape. Beyond the demographic aspects, assigning children to public schools is a complex task as it has to satisfy political and public demands, as well as to meet legal requirements. To help to the administrative services in performing this task we have developed a fuzzy-logic approach for children to school assignment.

Models for schools and schoolchildren location-allocation have been investigated in operational research ever since the 60's (Clarke and Surkis (1968), Koenigsberg (1968), Heckman and Taylor (1969)): see the reviews of Caro et al. (2004), as well as Castillo-Lopez and Lopez-Ospina (2015) for recent years. All the models seek to to allocate geography units to facilities, while considering spatial (journey to school distance treshold, closest assignment), logistic (capacity of schools) (Delmelle et al. (2014)), segregation and or economic constraints (Antunes and Peeters (2000)). Minimizing the average distance for given children locations and school capacities amounts to the celebrated optimal transportation problem (see e.g. Villani (2009)), possessing in general many equivalent solutions, computationally demanding for large networks.

Minimizing the social segregation at schools also possesses a multitude of solutions whenever the number of children locations exceeds the number of social categories, as in real situations.

The intrinsic degeneracy of the minimal distance and the minimal segregation problems prevents the existence of a unique and stable spatial partition, made of the union of school assignation basins. By contrast, introducing an additional origin-destination mutual information term, measuring the assignment *cohesiveness* and penalizing hard assignments, permits to make the solution unique. It furthermore allows the quick iterative computing of the minimal distance and the minimal segregation problems, as well as convex combinations of the latter, provided the penalization term is large enough.

The formalism and issues presently addressed bear obvious connections with iterative fitting, doubly constrained gravity flows, regularized optimal transportation, and model-based clustering with fixed mixture proportions. It involves in addition a presumably original term, the *social segregation at destination*, and appears to be *analytically tractable* (iterative determination of the unique solution, duality theory and sensitivity analysis). All the ingredients can be combined into an unique objective functional, the *free energy*, depending upon two freely adjustable parameters, the *temperature* T controlling the penalization of hard assignments, and the *mixing parameter* c, favoring social diversity at the destinations (section II). Toy and real applications (section III) illustrate the formalism together with its arguably numerous virtues, and the conclusion (section IV) emphasizes its connections to sensitivity analysis and robustness.

II FORMALISM

Formalizing the above considerations necessitates the introduction of three categorical variables, namely O = "child domicile" with n modalities, G = "school locations " with m modalities, and A = "social characteristics at origin" with p modalities.

2.1 Notations and definitions

Consider the $n \times m$ matrix $N = (n_{ig})$ counting the number of children living at i and assigned at school g. Row margins $n_{i\bullet}$ (where "$\bullet$" denotes the sum over the replaced index) count the number of children at i. Column margins $n_{\bullet g}$ yield the absolute capacity of school g.

Relative $n \times m$ *couplings* $P = (p_{ig})$ with $p_{ig} = n_{ig}/n_{\bullet\bullet} = p(i,g)$ constitute the joint probability distribution of O and G. Its margins are denoted by $f_i = n_{i\bullet}/n_{\bullet\bullet} = p(i)$ (origin weights) and $\rho_g = n_{\bullet g}/n_{\bullet\bullet} = p(g)$ (destination weights). The assignment *cohesiveness* is measured by the origin-destination mutual entropy, that is

$$I[P] := I(O:G) = \sum_{ig} p_{ig} \ln \frac{p_{ig}}{f_i \rho_g} = H(O) + H(G) - H(O,G) = H(G) - H(G|O) \quad (1)$$

where $H()$ denotes the corresponding entropies. By construction, $I[P] \geq 0$, with equality iff $p_{ig} = f_i \rho_g$, that is iff O and G are independent.

Let $D = (d_{ig})$ denote the $n \times m$ matrix of domicile to school pedestrian distances. The transportation cost or *energy* is the average journey to school distance:

$$U[P] := \sum_{ig} p_{ig} \, d_{ig} \quad (2)$$

Finally, consider a "social variable" A (such as nationality, ethnicity, social class or sex) with p modalities $a = 1, \ldots, p$, and let $Y = (y_{ia})$ denote the given $n \times p$ matrix specifying the proportion of social type a at origin i, i.e. $y_{ia} = p(a|i)$ with $y_{ia} \geq 0$ and $y_{i\bullet} = 1$. The joint distribution of social types and destinations is given by $Q = (q_{ag})$ with

$$q_{ag} = \sum_i y_{ia}\, p_{ig} = p(a, g) \qquad \text{with} \qquad q_{a\bullet} = \sum_i f_i y_{ia} =: r_a \qquad \text{and} \qquad q_{\bullet g} = \rho_g \ .$$

$r_a = p(a)$ is the overall proportion of type a. Let us measure the social segregation at destinations by the mutual information between A and G, that is

$$S[P] := I(A : G) = \sum_{ag} q_{ag} \ln \frac{q_{ag}}{r_a \rho_g} \tag{3}$$

By construction, $S[P] \geq 0$, with equality iff $q_{ag} = r_a \rho_g$, that is iff A and G are independent (absence of segregation at destinations).

2.2 Minimizing the free energy

The simultaneous control of the average distance, segregation and cohesiveness can be achieved by minimizing the *free energy*

$$F[P] := U[P] + c\, S[P] + T\, I[P] \tag{4}$$

among all couplings P with given margins f and ρ. Here $T \geq 0$ is a freely adjustable *temperature* parameter, and $c \geq 0$ a freely adjustable *mixing* parameter. Optimal transportation problem corresponds to $T = c = 0$. Setting $T > 0$ favors soft assignment, and setting $c > 0$ diminishes social segregation at destinations. $c < 0$ can also be adopted, *favoring social segregation*, provided $|c|$ is not too large: further analysis demonstrates that $F[P]$ is convex for $c \geq 0$ (and hence possesses an unique minimizer), but ceases to be convex for $c < -T$.

The minimization of $F[P]$ is a standard exercise in calculus. Setting to zero its first derivative (including the Lagrange multipliers term insuring that marginal constraints are satisfied) yields the identities

$$p_{ig} = f_i \rho_g \exp(-\beta\, d_{ig}) \exp(-\beta\, c\, \ell_{ig}) \exp(\beta(\lambda_i + \mu_g)) \tag{5a}$$

$$\exp(-\beta\lambda_i) = \textstyle\sum_g \rho_g \exp(\beta\mu_g) \exp(-\beta\, d_{ig}) \exp(-\beta\, c\, \ell_{ig}) \tag{5b}$$

$$\exp(-\beta\mu_g) = \textstyle\sum_i f_i \exp(\beta\lambda_i) \exp(-\beta\, d_{ig}) \exp(-\beta\, c\, \ell_{ig}) \tag{5c}$$

$$\text{where} \quad \ell_{ig}[P] := \sum_a y_{ia} \ln \frac{q_{ag}[P]}{r_a \rho_g} = \sum_a \frac{p(ia)}{p(i)} \ln \frac{p(ag)}{p(a)p(g)} \ . \tag{5d}$$

Here $\beta = 1/T$ is the inverse temperature or *coldness*, λ_i and μ_g the Lagrange multipliers insuring the origin, respectively destination constraints. The quantity ℓ_{ig} (possibly negative) behaves as a "social distance" in the expression $d_{ig} + c\,\ell_{ig}$, penalizing the assignment $i \to g$ whenever the dominant social characteristics a at i tend to be overrepresented at g.

Direct substitution shows the minimum free energy to be $F = \sum_i f_i \lambda_i + \sum_g \rho_g \mu_g$, an expression generalizing Kantorovitch duality theory of optimal transportation (e.g. Villani (2009)), and justifying the interpretation of the multipliers λ_i and μ_g as the unit embarkment cost at origin i, respectively disembarkment cost at destination g.

2.3 Iterative determination of the optimal coupling

Starting from some initial coupling such as $P^{(0)} = f\rho'$, and some destination multiplier value such as $\mu^{(0)} = 0$, a new coupling $P^{(1)}$ is obtained by successively applying (5d), (5b), (5c) and (5a). The process is then iterated. Convergence occurs provided the cohesiveness contribution is large enough compared to the contributions of both energy and social segregation. That is, the temperature T must be large enough to allow efficient regularization. Numerical experiments with `R` (2.15.2) for the toy network of section 3.2 shows that convergence occurs (say after ca. 1'000 iterations) provided both the following conditions hold:

i) $T \geq 0.003 \cdot d_c$, where d_c is the *Hausdorff distance* between the set (support) of sources $S = \{i|f_i > 0\}$ and the set of targets $T = \{g|\rho_g > 0\}$, that is $d_c = \max(d'_c, d''_c)$, where $d'_c{=}\max_i \min_g d_{ig}$ and $d''_c{=}\max_g \min_i d_{ig}$. Below that threshold, overflow occurs in (5b) or (5c).

ii) $T \geq |c|$, that is $c \in [-T, T]$. Outside that range, the iterative process may exhibit periodic or non-converging behavior (especially for small values of T), betraying instability of the fixed point (yet existing and unique for $T > 0$ and $c \geq 0$ since $F[P]$ is convex).

2.4 Boundaries and placement error

High temperature favors low cohesiveness, that is soft memberships (of children into schools) $z_{ig} = p_{ig}/f_i = p(g|i)$. The conditional entropy $H(G|i) = -\sum_g z_{ig} \ln z_{ig} \geq 0$ measures the uncertainty in the assignment of i. High values of $H(G|i)$ characterize origins i lying at the *boundary* of two or more assignment basins, and help selecting domicile candidates for school reassignments, if needed (e.g. changes in f, ρ or Y from year to year).

This being said, even origins far from the assignment boundaries obey $H(G|i) > 0$ for $T > 0$ in view of the softness of Z. Attributing a given origin to different schools is not always possible nor desirable in real situations, so Z might have to be *hardened* into $Z^\star = (z^\star_{ig})$, where $z^\star_{ig} = 1$ if $g = \arg\max_h z_{ih}$, and $z^\star_{ig} = 0$ otherwise (ties are resolved at random). In general, $Z^\star$ does not satisfy the capacity constraint, that is the quantity $\rho^\star_g = \sum_i f_i z^\star_{ig}$ generally differs from ρ_g: for f and ρ given, equation $\rho = fZ'$ possesses generally no hard solution Z. The relative placement error $\delta = \sum_g |\rho_g - \rho^\star_g|$ is expected to increase with T (figure 1). The absolute placement error $N\delta$ (where N is the total number of children) counts the number of misattributed children (excess or deficit) by the hardened assignment.

III ILLUSTRATIONS AND APPLICATIONS

3.1 Example A

$n = 600$ children of the same kind (no segregation considerations here) are uniformly placed on a 25×24 grid, and have to be assigned to $m = 6$ schools of identical capacity of 100 each, so that the average city-block distance (L_1 metric) is minimal (1). As expected, the fuzziness of the boundaries as well as the placement error diminishes with the coldness $\beta = 1/T$.

3.2 Example B

Consider the random placement, without overlapping, of $n = 40$ origins and $m = 5$ destinations, on a 13×13 grid endowed with the city-block distance, with uniform origin weights $f_i = 1/n$ and uniform destination weights $\rho_g = 1/m$. In addition, we focus on gender segregation ($p = 2$), and generate a gender gradient across the origins, such that the proportion of girls is maximum in the south-west corner, and minimum in the north-east corner. Figure (2) depicts the hardened memberships $Z^\star$ for each origin i, and also the proportion of girls versus boys at domiciles and at schools.

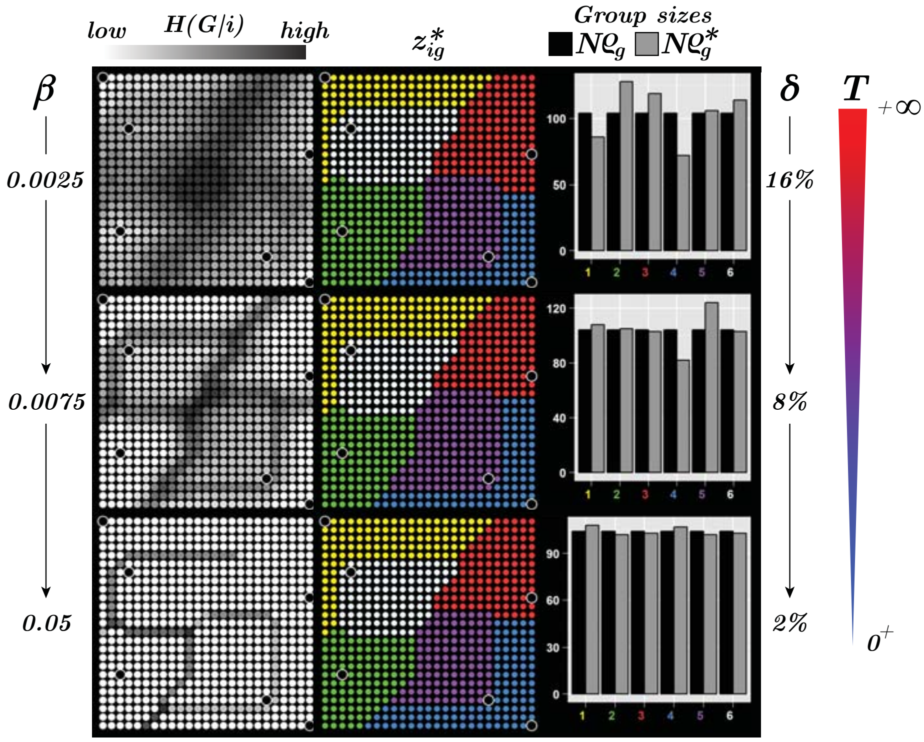

Figure 1: Example A: entropic boundaries $H(G|i)$ between school assignment basins, home-to-school assignment based upon the hardened membership $z^{\star}_{ig}$ (section 2.4), group sizes and resulting placement error δ, for decreasing values of the temperatures $T = 1/\beta$.

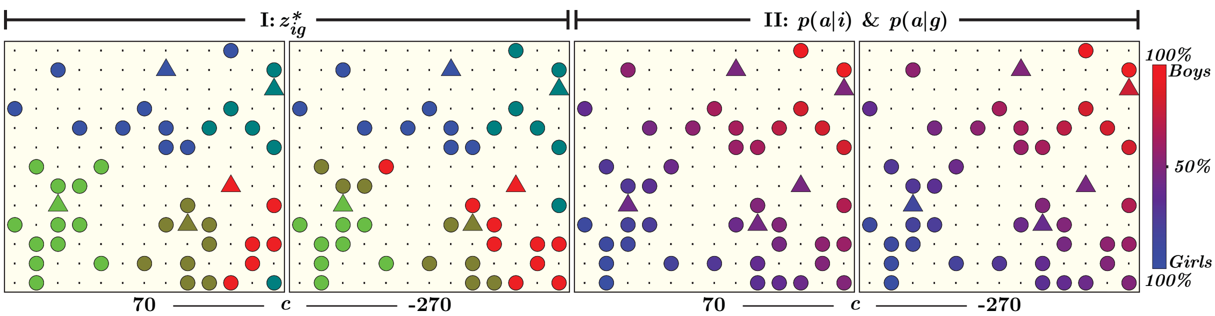

Figure 2: Example B. I: Hardened membership $z^{\star}_{ig}$ (I), II: the proportion of girls versus boys at origins $p(a|i)$ (circles) and at destinations $p(a|g)$ (triangles). With $d_c = 7$, $T = 16$ and while penalizing gender segregation at the destinations ($c = 70 > 0$) or, on the opposite, favoring it ($c = -270 < 0$).

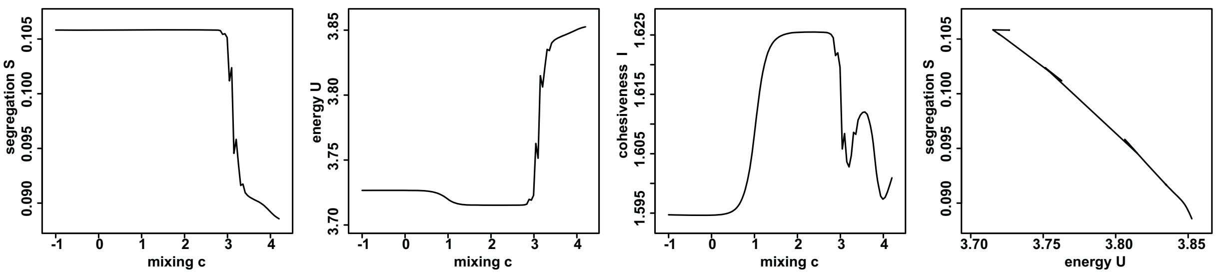

Figure 3: Example B: at low temperature $T = 0.02$, the iterative solution converges for $c \in [-1, 4.2]$. Mixings $c \in [0, 3]$ increase the cohesiveness I but leave U and S unchanged. Values $c > 3$ produce a sharp decrease in segregation S, and a sharp increase in energy U, depicting an effective anti-segregation regime together with an additional average distance cost.

3.3 Children-to-school constrained assignment in Lausanne

In the optimal transportation limit, that is in absence of social segregation considerations (i.e. $c = 0$), and for low temperatures $T \to 0^+$, the procedure has been shown (Guex et al. (2016), Emmanouilidis (2016)) to permit mapping of school attendance areas together with their fuzzy boundaries. As shown in Figure 4 (left), large values of the (spatially smoothed[1]) assignment entropy $H(G|i)$ (section 2.4) highlight fuzzy boundaries between basins, where the origin-to-destination attribution is most uncertain, and where the potential for school organisation and planning is most promising (e.g. balancing class sizes, or splitting up neighborhoods between two or more schools in order to promote social mixing). Furthermore, groups of attendance areas can be aggregated to design a new school district pattern. Mapping Lagrangian multipliers provides an overview of embarkment costs λ_i and disembarkment costs μ_g (Figure 4, right). Such a visualization can be especially useful for the comparison and evaluation of various potential scenarios involving changes of school capacities or new schooling infrastructures.

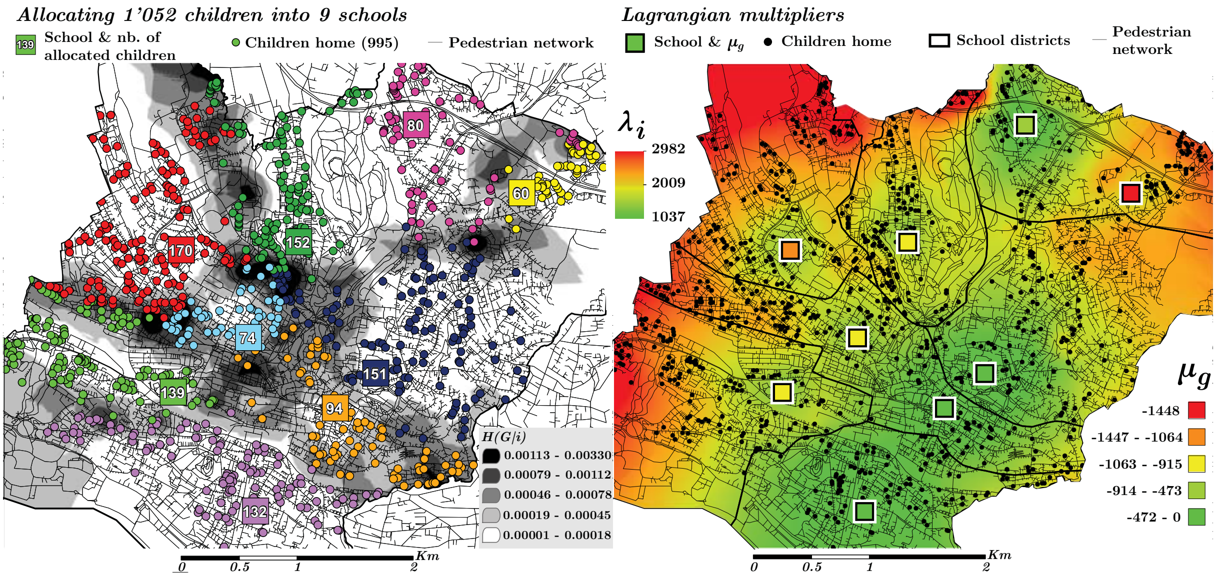

Figure 4: Resulting allocation plan for children entering secondary school in summer 2016 ($\beta = 0.2$). Left: hard assignment $Z^\star$ generates fairly compact attendance areas around schools. Children located in dark areas (high entropy $H(G|i)$) are more likely to be assigned to two or more schools, than the ones living in white areas. Right: the lack of schools in the western and northern outskirts of the city generates high (red orange) values of λ_i (spatially smoothed large embarkment costs). By contrast, children living in the center and in the south benefit from shorter journeys to school (green, yellow). The top-right school is too small to enroll all the surrounding children, which results in a large disembarkment cost μ_g.

3.3.1 Assignment with gender segregation constraint

Running the analysis on a similar dataset with adding the mixing parameter c leads to the results exposed in Figure 5. We here consider the assignment of 973 children into 9 schools. Cases I and II relates respectively the results with high ($c = -700$) and low ($c = 300$) gender segregation, while case III is free of segregation constraint ($c = 0$). Despite the almost perfect balance between the number of boys (486) and girls (487), their repartition within the city is locally heterogeneous (background maps of line B). Allocating children with high segregation ($c = -700$) generates almost five unisex schools (maps I-B & C). Four of them are located in

[1] in this paper, spatial interpolation is systematically performed through krigging (ArcGIS settings: type spherical, variable search radius, number of points:10).

the city-center while the fifth, on the top, is surrounded by a majority of girls. Highest values of c promotes, on the opposite, gender mix within schools (II-B & C). Results obtained with $c = 0$ are almost the same ((III-B & C), showing a good gender balance and suggesting that proportion of boys and girls around schools are almost equal. Accordingly, gender segregation here involves spatial mixing of overlapping hard assignments (I-A), while gender mixing creates a spatial partition of homogeneous connected groups (II & III-A).
Although the range of the assignment uncertainty ($H(G|i)$, line D) tends to narrow as c decreases, highest values characterize the same city parts in the three maps. The Lagrangian multipliers (line E) exhibit similar trends in cases I and II, where children of the west, north and far east suburbs have the longest journey to school (highest λ_i). With $c = 300$ (II-E) embarkment costs in these areas are more smoothed. The disembarkment costs μ_g vary little between the three cases, suggesting that gender segregation has, for this dataset, little effect on the former.

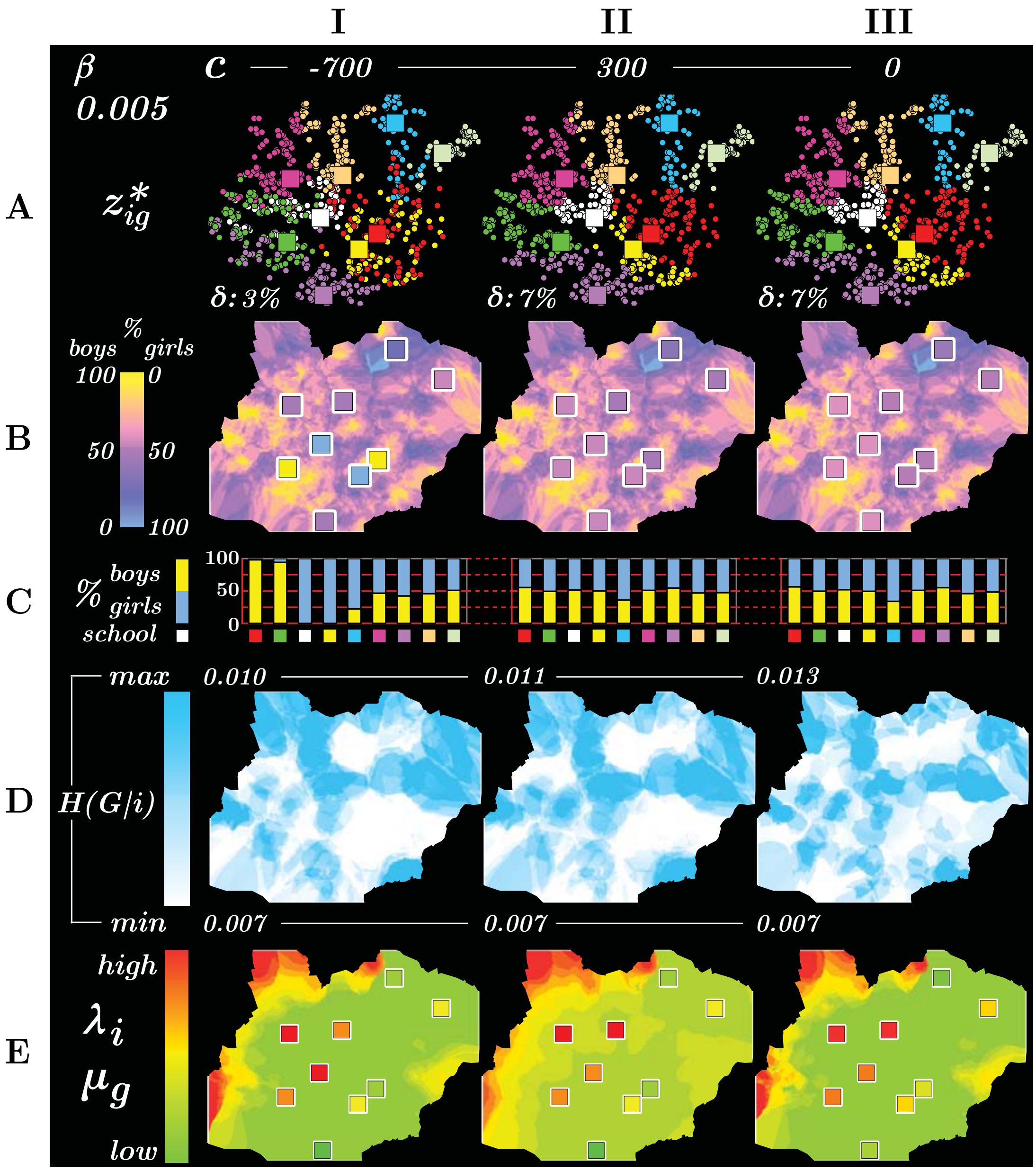

Figure 5: Children to school assignment with gender segregation constraint. A large negative value of c in (I) produces segregative allocation plan tending towards unisex schools. On the opposite, a large positive value of c promotes gender mixing within the schools.

 [ISBN: 978-2-9105-4510-5]

IV CONCLUSION

Introducing some amount of randomness and uncertainty in an otherwise deterministic problem (the optimal transportation) results in a unique solution, easily and quickly computable, and *robust* with respect to small variations of child locations and populations, or school locations and capacities: once more, the regularizing virtues of the origin-destination mutual entropy are confirmed. Regularization permits in addition the effective identification and visualization of borders, of costs, and of alternative near-optimal solutions: those benefits are essential (and much appreciated) for adapting assignments to changes (turn-over of children at the origins, modifications of the pedestrian network) and for future planning.

As demonstrated in the present paper, this approach can be extended to handle in addition the issue of *social segregation*, by the introduction of a new *mixing parameter* controlling for the segregation. The resulting thermodynamic-like formalism permits to perform tractable *sensitivity analyses* of various kinds, assessing the amount of change in the quantities of interest induced by a change in the network geometry, origin or destination distributions, or by a change in the temperature or mixing parameters. In particular, the distance versus segregation trade-off can be made fully explicit (figure 3) - a crucial prerequisite for informed political decisions.

Instead of using pedestrian distances, the algorithm can be also be fed by alternative transportation costs, such as the total displacement time on uni- or multimodal transportation unoriented networks. Also, beside gender mixing, the segregation issues related to child age or ethnic origin can be tackled by the same approach, to be addressed in a future work. Among further possible developments, the question of dealing with flexible school penalized capacities (instead of with fixed capacities) seems worth investigating, as is the issue of imposing a minimal social mixing at each school, rather than on average only.

References

Antunes A., Peeters D. (2000). A dynamic optimization model for school network planning. *Socio-Economic Planning Sciences 34*(2), 101–120.

Caro F., Shirabe T., Guignard M., Weintraub A. (2004). School redistricting: Embedding gis tools with integer programming. *Journal of the Operational Research Society 55*(8), 836–849.

Castillo-Lopez I., Lopez-Ospina H. A. (2015). School location and capacity modification considering the existence of externalities in students school choice. *Computers & Industrial Engineering 80*, 284–294.

Clarke S., Surkis J. (1968). An operations research approach to racial desegregation of school systems. *Socio-Economic Planning Sciences 1*(3), 259–272.

Delmelle E. M., Thill J.-C., Peeters D., Thomas I. (2014). A multi-period capacitated school location problem with modular equipment and closest assignment considerations. *Journal of Geographical Systems 16*(3), 263–286.

Emmanouilidis T. (2016, 02). *Analyse de réseau piéton et gestion scolaire: sinuosité, centricité et transport optimal régularisé*. Ph. D. thesis, Université de Lausanne, faculté des Géosciences et de l'Environnement.

Guex G., Emmanouilidis T., Bavaud F. (2016). Transportation clustering: a regularized version of the optimal transportation problem. *Submitted for publication*.

Heckman L. B., Taylor H. M. (1969). School rezoning to achieve racial balance: a linear programming approach. *Socio-Economic Planning Sciences 3*(2), 127–133.

Koenigsberg E. (1968). Mathematical analysis applied to school attendance areas. *Socio-Economic Planning Sciences 1*(4), 465–475.

Villani C. (2009). *Optimal Transport, Old and New*. Springer.