Mazzitello, Karina I.; Candia, Julián

Diffusion-Based Density-Equalizing Maps: an Interdisciplinary Approach to Visualizing Homicide Rates and Other Georeferenced Statistical Data

GENERAL AND APPLIED PHYSICS

Diffusion-Based Density-Equalizing Maps: an Interdisciplinary Approach to Visualizing Homicide Rates and Other Georeferenced Statistical Data

Karina I. Mazzitello · Julián Candia

Abstract In every country, public and private agencies allocate extensive funding to collect large-scale statistical data, which in turn are studied and analyzed in order to determine local, regional, national, and international policies regarding all aspects relevant to the welfare of society. One important aspect of that process is the visualization of statistical data with embedded geographical information, which most often relies on archaic methods such as maps colored according to graded scales. In this work, we apply nonstandard visualization techniques based on physical principles. We illustrate the method with recent statistics on homicide rates in Brazil and their correlation to other publicly available data. This physics-based approach provides a novel tool that can be used by interdisciplinary teams investigating statistics and model projections in a variety of fields such as economics and gross domestic product research, public health and epidemiology, sociodemographics, political science, business and marketing, and many others.

arXiv:1206.4777

K. I. Mazzitello
CONICET and Departamento de Física, Facultad de Ingeniería, Universidad Nacional de Mar del Plata, Mar del Plata, Argentina
e-mail: kmazzite@mdp.edu.ar

J. Candia (✉)
CONICET and Instituto de Física de Líquidos y Sistemas Biológicos, Universidad Nacional de La Plata, La Plata, Argentina
e-mail: juliancandia@gmail.com

1 Introduction

Nowadays, interdisciplinary efforts in science directed towards the most acute real-world problems of society are being increasingly encouraged and stimulated. And, rightly so, this task is even more urgent in the Latin American region, which suffers from some of the highest rates of crime and violence in the world [1] and their devastating effects on citizen security and human rights [2], as well as from other endemic issues such as widespread illiteracy, sharp socioeconomic disparities, undernourishment, etc. In this context, the statistical physics community is in an excellent position to contribute invaluable new tools and insight to areas in which statistical analysis is performed by long-standing, more traditional means.

One of the important aspects of dealing with statistical information is data visualization. Indeed, as we enter the age of the petabyte, where we routinely capture, warehouse, and analyze massive amounts of data [3], pioneering methods of visualization are needed to change the way we understand and disseminate science. Some examples of new visualization techniques recently developed are state-of-the-art animations to see the quantum world, accurate maps of the universe based on the Sloan Digital Sky Survey covering large areas of the sky, and complex network representations that facilitate the recognition of patterns in multidimensional complex systems [4–6].

In the field of applied statistics, traditional visualization tools such as bar and pie charts, regression plots,

and box plots are still the most widely used methods. Statistical data with embedded geographical information (for instance, census data) are either organized in tables or shown by means of standard maps with color codes by region. Such maps, however, can be misleading. Often times, statistical measures are correlated with the population density and other indicators, but those correlations remain hidden from the representation. For instance, a plot of disease incidence will inevitably show a high incidence in cities and low incidence in rural areas solely because more people live in cities. An alternative representation is to plot the incidence per capita, which solves this problem at the cost of discarding all information about where most of the cases are occurring. In order to meaningfully visualize statistical data, bringing together relevant statistical indicators and the embedded geographical information, a different approach is needed.

Over the last few decades, some visualization work has focused on creating so-called *cartograms*, in which geographic regions are deformed and rescaled in proportion to their population or other statistical indicators. Several of these efforts were inspired by physics. For instance, in the approach by Gusein-Zade and Tikunov [7], a continuous displacement field is used to expand high population areas by means of a repulsive force. Other methods by Appel et al. [8] and by Dorling [9] are based on cellular automata, in which local interaction rules allow boundaries to shrink or expand locally until equilibrium is reached. Another procedure proposed by Kocmoud [10] relies on a mechanical model of masses and springs, which is constrained in order to maintain certain topographic features. Despite these efforts, cartograms built following those methods proved problematic in a number of ways, leading to highly distorted shapes, overlapping regions, or strong dependences on the choice of coordinate axes. More recently, Gastner and Newman [11] proposed a new method for producing cartograms based on diffusion equations, which resolves the shortcomings experienced with previous techniques. Starting with an inhomogeneously distributed statistical quantity (e.g., the population density), diffusion equations are numerically integrated with appropriate boundary conditions. By allowing the diffusion process to evolve until a homogeneous equilibrium state is reached, the displacements can be reinterpreted to generate a cartogram.

Within this context, in this work we implement the diffusion-based method to create a variety of density-equalizing maps of the 27 Brazilian federative units. With this approach we generate a series of cartograms that show different ways of displaying population,

homicide rates, and a combined index of education, health, and economic development. Depending on the way the cartograms are built and how the statistics are chosen, different perspectives can be gained from the data. Also, cross-correlations between statistical quantities and their dynamical evolution can be represented in more effective ways than standard visual methods. The aim of this paper is to show how a novel physics-based visualization tool can be applied to real-world data of great cross-disciplinary impact, as well as to discuss the benefits that can be drawn from the new approach compared to the visual methods usually employed by statisticians of public and private agencies.

The rest of this paper is organized as follows: In Section 2, we present details on the diffusion-based method to generate density-equalizing maps. Section 3 is devoted to the presentation and discussion of the cartograms, while our conclusions are finally stated in Section 4.

2 The Diffusion-Based Cartogram Method

In this section, we summarize the diffusion-based method for generating density-equalizing maps [11]. Let us consider $\rho_0(\vec{r})$, the spatial density distribution of the inhomogeneous statistical indicator that we want to represent by means of a rescaled map. We will call $\rho_0(\vec{r})$ the *population density*, since the population distribution is one of the indicators one may be interested in. Notice, however, that the method can be applied to any other statistical quantity distributed inhomogeneously in space, as will be shown in the examples of Section 3. The basic idea is to rescale the map's regions to become proportional to their population so that the population density on the rescaled map is homogeneously distributed. If the total area of the map is preserved, the homogeneous population density after rescaling will be equal to the mean population density before rescaling, $\bar{\rho}_0$.

From a mathematical point of view, any total-area-preserving map rescaling operation corresponds to a coordinate transformation $\vec{r} \to \vec{r}'$ under the condition that the determinant of the corresponding Jacobian matrix equal the density ratio $\rho_0(\vec{r})/\bar{\rho}_0$. Although this condition guarantees that the total area of the map remains constant, it does not provide specific information about the cartogram projection. In order to determine a unique projection, we resort to the physical process of diffusion by considering the population density as a time-dependent scalar field, $\rho(\vec{r}, t)$, that satisfies the initial condition $\rho(\vec{r}, t = 0) \equiv \rho_0(\vec{r})$. From

a macroscopic perspective, the current density, related to time-dependent velocity and density fields through

$$\vec{J} = \vec{v}(\vec{r}, t)\rho(\vec{r}, t) , \tag{1}$$

is directed along the gradient of the density field:

$$\vec{J} = -D\nabla\rho , \tag{2}$$

where D is the diffusion constant. Without loss of generality, we can conveniently select time and length units in order to set $D \equiv 1$. These equations, together with the equation for the conservation of mass,

$$\nabla \cdot \vec{J} + \frac{\partial\rho}{\partial t} = 0 , \tag{3}$$

lead to the well-known diffusion equation:

$$\nabla^2\rho - \frac{\partial\rho}{\partial t} = 0 . \tag{4}$$

We enclose the area of interest within a larger box, whose initial population density matches the mean population density $\bar{\rho}_0$. The role of this box is merely to act as a large container and avoid finite-size effects due to the system's boundaries. By solving this diffusion equation with the initial condition $\rho(\vec{r}, t = 0) \equiv \rho_0(\vec{r})$ and Neumann boundary conditions such that no flow occurs through the boundaries of the container, the density field can be determined at all times. Then, the velocity field follows as

$$\vec{v}(\vec{r}, t) = -\frac{\nabla\rho(\vec{r}, t)}{\rho(\vec{r}, t)}, \tag{5}$$

and from the latter, the cumulative displacement of any point in the map can be calculated by solving the kinematics, i.e.,

$$\vec{r}(t) = \vec{r}(0) + \int_0^t \vec{v}(\vec{r}, t')dt' . \tag{6}$$

The cartogram is thus derived by moving all region boundaries on the map in such a way that the net flow passing through them is zero. In the large-t limit, the population is uniform everywhere within the box and the regions become rescaled to sizes proportional to their initial population densities.

As a simple example to illustrate the procedure, consider the system shown in Fig. 1 by the black dashed lines. It consists of two regions of equal size: region A on the left has population density $\rho_{0,A} = 1$ (in arbitrary units) and region B on the right has $\rho_{0,B} = 2$. We embed both regions within a large enclosure filled with the average density $\bar{\rho}_0 = 1.5$. Letting the concentrations diffuse, the system evolves towards homogeneity. Once equilibrium is attained, the region boundaries have been reshaped as shown by the red solid lines. The rescaled map shows that region B is twice as large
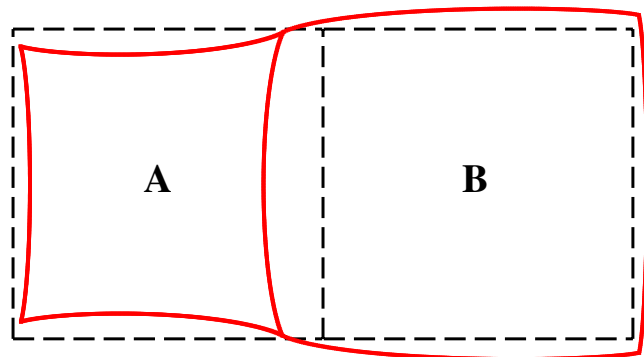

Fig. 1 (Color online) Example showing the initial areas (enclosed by *black dashed lines*) and final areas (enclosed by *red solid lines*) when the population density of region B is twice as large as that of region A

as region A, as expected from the initial population density ratio. That is, the ratio of final areas when the diffusion process reaches equilibrium, Area(B)/Area(A) = 2, agrees with the ratio of initial densities, $\rho_{0,B}/\rho_{0,A} = 2$. In this way, the rescaled map has merged the geographical information with the statistical information: while regions A and B are still recognizable, the rescaled representation also tells us that B is twice as large as A in the sense of the chosen statistical measure. The next section is devoted to actual examples showing different ways to implement these ideas.

3 Density-Equalizing Cartograms of Brazil: Results and Discussion

In this section, we implement the diffusion-based method to generate density-equalizing cartograms of Brazilian states. By using recent, publicly released statistical data provided by Brazilian government agencies, we illustrate the method and discuss different scenarios to apply this visualization technique.

Being one of the major challenges facing Brazil and, on a larger context, the rest of Latin American countries as well, the geographical distribution of homicide statistics, their cross-correlation with other statistical measures, and their time evolution were the topics we focused on. To that end, we examined data on nationwide violent crime statistics very recently released in a joint report by the Brazilian Ministry of Justice and the Instituto Sangari [12]. Demographic data were obtained from the IBGE [13], Brazil's chief government agency responsible for statistical and geographical information, whose Portuguese acronym stands for "Brazilian Institute of Geography and Statistics." It

is interesting to point out, in passing, that the tool presented in this work integrates *geography* and *statistics* into a unified visual framework, thus naturally reconciling the two main subject areas of this agency. Furthermore, we used the Índice FIRJAN de Desenvolvimento Municipal (IFDM) index from the so-called "FIRJAN System" [14], which measures the combined education, health, and economic development at the municipal level based on official data released by the Brazilian Ministries of Education, Health, and Labor.

Homicide rates by state are usually depicted as in Fig. 2, in which the shapes and areas of the states are faithful geographical representations, while the different homicide rates are shown by means of color or grayscale shadings. This accurate geographical information comes at a price: from this representation, we do not know which states or regions are the ones where most homicides occur. Moreover, standing out in this map are the states with large surface areas, which might not be relevant neither by having particularly high homicide rates nor high absolute number of homicide counts. A good example in this regard is offered by the largest state, Amazonas, which ranks 17th among the 27 federative units (i.e., the 26 states plus the Federal District) in the list of states with highest homicide rates and also 17th in the list of states with highest absolute homicide counts (using statistics corresponding to 2008 [12]). In contrast, the top-ranking state in homicide rates, Alagoas, appears barely noticeable due to its very small surface area.

Fig. 3 Cartogram showing areas proportional to each state's population and *grayscale shading* corresponding to homicide rates per 100,000 inhabitants in 2008. Notice the shrinkage of the northern and center-west regions in contrast with the expansion of the Federal District and the southern, southeastern, and northeastern regions. By paying the penalty of a highly distorted geography, we integrate the information of homicide rates (*shading*) with population size (*areas*)

In order to put in evidence the most populated states, we generate the cartogram of Fig. 3 by assuming $\rho_0(\vec{r}) \propto N_i/A_i \; \forall \vec{r} \in A_i$, where N_i is the population and A_i the surface area of the ith state. The density-equalizing diffusion-based method thus leads to a representation in which the area of each state is in proportion to its population. As in Fig. 2, the grayscale shading shows each state's homicide rates per 100,000 inhabitants in 2008. Since the absolute number of homicides is obtained from the product of homicide rate times the population size, this representation arguably displays all three statistics (i.e., population, homicide rate, and homicide counts). For instance, large dark states in the northeastern region are emphasized, as well as Espírito Santo and Rio de Janeiro, as most affected by violent crime both in relative and absolute terms. Although having one of the lowest homicide rates in Brazil, São Paulo stands out at the center of the cartogram due to its being the most populated state in the country. Naturally, the cartogram approach trades off geographical accuracy in exchange for incorporating additional information. In the following figures, we discuss alternative ways to implement such trade-offs.

Figure 4 shows a compromise between the geographically accurate, standard map of Fig. 2 and the highly

Fig. 2 Standard visualization of georeferenced statistical data: the shape and area of each Brazilian state is a faithful geographical representation, while the *grayscale shading* shows homicide rates per 100,000 inhabitants using 2008 data from Ref. [12]. The highest homicide rate is 60.3 (Alagoas, in *black*) and the lowest one is 12.4 (Piauí, in *white*)

Fig. 4 Cartogram showing the change of areas (relative to its geographically faithful size) proportional to each state's population and *grayscale shading* corresponding to homicide rates per 100,000 inhabitants in 2008. This cartogram is a compromise between the standard representation of Fig. 2 and the highly distorted cartogram of Fig. 3

distorted cartogram of Fig. 3. In this cartogram, which is obtained from assuming $\rho_0(\vec{r}) \propto N_i \; \forall \vec{r} \in A_i$, the *change* in area (relative to the geographically accurate representation) is proportional to each state's population. The grayscale shading, as before, corresponds to homicide rates per 100,000 inhabitants in 2008. Therefore, each state's area on this cartogram conveys some information on the geographical surface area, which was completely missing from the cartogram of Fig. 3. Moreover, the shapes are less distorted than in the previous cartogram. It should be pointed out, however, that since this cartogram integrates geographical and statistical information in a nontrivial manner, its interpretation is rendered more difficult. For instance, the similarity of two states depicted with similar sizes on the cartogram could be due to different combinations of population and geographical surface area of those states. In Fig. 4, São Paulo and Minas Gerais appear represented with similar sizes, which is due to the fact that São Paulo's population is roughly twice as large, while its surface area is roughly half as large, as those of Minas Gerais.

In Fig. 5, the areas are proportional to each state's homicide rates per 100,000 inhabitants in 2008, while the logarithmic grayscale shading reflects each state's population. That is, this figure reverses the roles of the two statistics shown earlier in Fig. 3. This cartogram is obtained by assuming $\rho_0(\vec{r}) \propto R_i/A_i \; \forall \vec{r} \in A_i$, where R_i is the homicide rate of the ith state. The cartograms in Figs. 3 and 5 contain the same information, yet their appearance is very different. One could argue that the focus of attention is first guided towards the large states and, therefore, size takes precedence over shade. Under that assumption, Fig. 3 places more emphasis on population (by noticing first the most populated states and only later their corresponding homicide rates), while conversely, Fig. 5 emphasizes homicide rates. Although the preferential focus of attention (size, shape, or color) might depend on the subject, it is interesting to point out that, as exemplified here, the same data can be visualized in different ways that imply a different emphasis on the represented quantities.

Uncovering correlations between crime statistics and other demographic, social, and economic variables is an active focus of research aiming at understanding the conditions that foster criminal behavior. For instance, Gould et al. found that low wages and unemployment make less-educated men more likely to turn to crime [15]. The impact of unemployment on crime has also been studied very recently using country-level statistical data from Europe [16]. Other recent studies

Fig. 5 Cartogram showing areas proportional to each state's homicide rates per 100,000 inhabitants in 2008 and logarithmic *grayscale shading* corresponding to each state's population. In this case, the cartogram tends to enhance the appearance of states and regions with high crime rates, as opposed to those with high population (compare with Fig. 3). For instance, Roraima and Amapá in the northern region stand out in this cartogram due to their significant crime rates but go unnoticed in Fig. 3 due to their small population

Fig. 6 Scatterplot showing homicide rates and the IFDM index (both corresponding to year 2008). The *solid line* is the best linear fit to the data
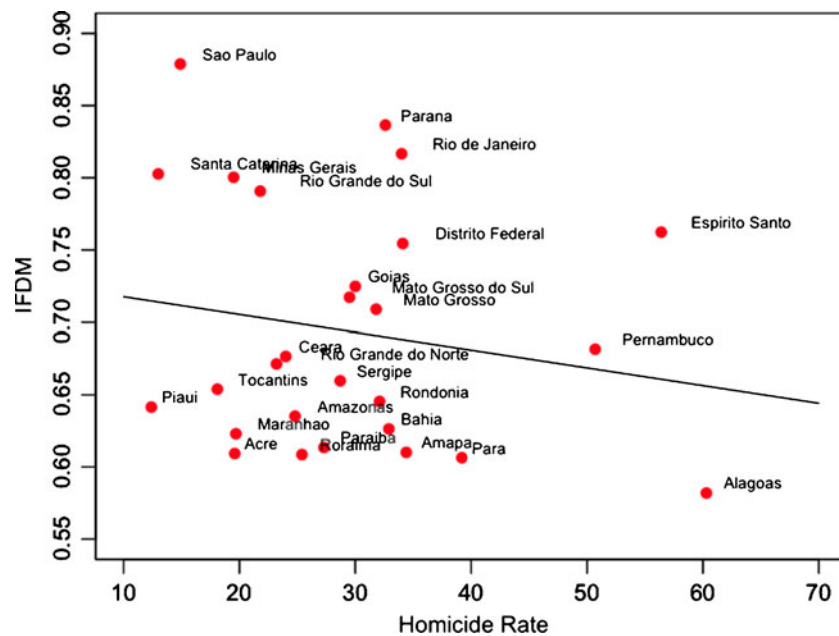
on crime-rate regressions are, e.g., Refs. [17–19] and references therein. Following our case study of state-level homicide rates in Brazil, let us now illustrate how correlations can be put in evidence by using diffusion-based cartograms.

As anticipated above, we used the IFDM index calculated from the so-called FIRJAN System [14], which measures the combined education, health, and economic development at the municipal level based on official data released by the Brazilian Ministries of Education, Health, and Labor. Since the IFDM index is a measure of welfare and prosperity, it is reasonable to expect states with higher IFDM scores to be less affected by violent crimes and vice versa. Indeed, the scatterplot in Fig. 6 hints to a plausible correlation between the two statistics, as evidenced by a best linear fit with negative slope (shown by the solid line) and a Pearson's correlation coefficient equal to -0.17. Notice, however, that the p value for the null correlation hypothesis is 0.385, and therefore, from a strict statistical perspective, we cannot rule out the absence of correlations. In the context of this paper, our choice of the IFDM index is only guided by the purpose of illustration of the cartogram method.

Figure 7 shows a cartogram that, as Fig. 5, displays the area of each state proportional to each state's homicide rates per 100,000 inhabitants in 2008. The grayscale shading represents each state's IFDM index (where darker states have lower IFDM scores). On the one hand, this cartogram puts in evidence the high geographical correlation of the IFDM index, with a dark fringe crossing the northern and northeastern regions,

while the states become increasingly lighter towards the south. This kind of geographical correlation, immediately apparent in the cartogram representation, is more

Fig. 7 Cartogram showing areas proportional to each state's homicide rates per 100,000 inhabitants in 2008 and *grayscale shading* corresponding to each state's IFDM index (where *darker states* have lower IFDM scores). The cartogram puts in evidence the high geographical correlation of the IFDM index, with a *dark fringe* crossing the northern and northeastern regions, while the states become increasingly lighter towards the south. Moreover, the cartogram captures plausible correlations between the IFDM index and crime rates, as for instance the *large, dark* northeastern regions (indicating low development and high crime rates) and, inversely, the comparatively *smaller and lighter* southern states

difficult to spot in the scatter plot of Fig. 6. On the other hand, moreover, the cartogram conveys plausible correlations with crime rates, as for instance the large, dark northeastern regions (indicating low development and high crime rates) and, inversely, the comparatively smaller and lighter southern states.

Finally, another interesting application of the diffusion-based cartogram representation is to capture the dynamical evolution of statistical quantities. By using annual data on homicide rates from 1998 to 2008, we generated additional frames by linear interpolation and created a movie showing substantial changes experienced over time (supplied as Electronic Supplementary Material to this paper). As opposed to analogous movies based on the usual colored or grayscale representation (Fig. 2), cartogram movies display changing features such as shape and size and therefore are particularly suitable for enhancing the visualization of dynamical observables. This technique can certainly improve the communication and outreach of statistical results to both specialized and nonspecialized audiences.

4 Conclusions and Outlook

Geography and statistics are ubiquitously intertwined in data collected to assess the levels of health, socioeconomic development, literacy and education, crime, and many other key indicators of the welfare of society. However, georeferenced data are most often represented by traditional visualization methods that do a poor job in integrating these two key aspects. In this work, we implemented a physics-based visualization approach to integrate geography and statistics into a unified framework. By trading off geographical accuracy in exchange for additional information, we explored different alternative scenarios that emphasize different views on the data. We believe that these novel techniques offer new avenues to communicate effectively and reach out to specialized and nonspecialized audiences in a wide variety of applied statistics fields, including economics and gross domestic product research, public health and epidemiology, sociodemographics, political science, business and marketing, and many others.

Although we illustrated the diffusion-based approach with Brazilian georeferenced statistical data at the federative unit level, the method can be extended to integrate different geographical and geopolitical scales, from local (e.g., town districts) to global (e.g., regions within a country and even relations among countries in the international context). Naturally, in order to fully capture the inherent complexities of georeferenced data across multiple dimensions, other methodologies are needed to complement the approach presented here. To that end, methods based on multilevel Voronoi tessellation [20] and multilayered complex network analysis and modeling [21, 22] offer promising avenues of research.

Besides the specific aim of improving the visualization of statistics, we hope that, in a broader context, this work will serve to illustrate possible cross-disciplinary endeavors addressing key local and regional issues. Indeed, we stress the importance of two aspects, namely, (1) recognizing the need for interdisciplinary work in many areas of applied science and the role physicists can play in providing new analysis methods and (2) recognizing the need to address key local and regional issues (e.g., within the Latin American context) in those areas where basic and applied research can find a common ground. Hopefully, this paper will contribute to those goals and stimulate further work.

References

1. L. Dammert, E. Alda, F. Ruz, *Desafíos de la seguridad ciudadana en Iberoamérica* (FLACSO, Chile, 2008)
2. J. Faroppa Fontana, Report on citizen security and human rights. (Inter-American Commision on Human Rights, Organization of American States, OEA/Ser.L/V/II. Doc. 57, 2009), www.cidh.org. Accessed 19 Apr 2012
3. C. Anderson, The end of theory: the data deluge makes the scientific method obsolete. Wired **16**(7) (2008)
4. B. Sanders, Seeing the quantum world. Physics World **21**(12)(2008)
5. M. Subba Rao, M. Aragon-Calvo, A picture of the cosmos. Physics World **21**(12) (2008)
6. C. Hidalgo, Thinking outside the cube. Physics World **21**(12) (2008)
7. S.M. Gusein-Zade, V.S. Tikunov, A new technique for constructing continuous cartogram. Cartogr. Geogr. Inf. Syst. **20**, 167 (1993)
8. A. Appel, C.J. Evangelisti, A.J. Stein, Animating quantitative maps with cellular automata. IBM Tech. Dis. Bull. **26**, 953 (1983)
9. D. Dorling, *Area Cartograms: their use and creation, concepts and techniques in modern geography (CATMOG)*, No. 59 (Univ. of East Anglia, Norwich, U.K., 1996)
10. C.J. Kocmoud, Constructing continuous cartograms: a constraint-based approach. Master's Thesis, Texas A&M Visualization Laboratory, Texas A&M University, College Station, 1997
11. M.T. Gastner, M.E.J. Newman, Diffusion-based method for producing density-equalizing maps. Proc. Nat. Acad. Sci. **101**, 7499 (2004)
12. J.J. Waiselfisz, Mapa da Violência 2011, Os Jovens do Brasil (Ministério da Justiça and Instituto Sangari, 2011)

13. IBGE's: http://www.ibge.gov.br/home. Accessed 19 Apr 2012
14. FIRJAN System's: http://www.firjan.org.br/ifdm/. Accessed 19 Apr 2012
15. E.D. Gould, B.A. Weinberg, D.B. Mustard, Crime rates and local labor market opportunities in the United States: 1979–1997. Rev. Econ. Stat. **84**, 45 (2002)
16. D.T. Altindag, Crime and unemployment: evidence from Europe. Int. Rev. Law Econ. **32**(1), 145–157 (2012)
17. C.E. Moody, T.B. Marvell, On the choice of control variables in the crime equation. Oxf. Bull. Econ. Stat. **72**, 696 (2010)
18. A. Cotte Poveda, Economic development, inequality and poverty: an analysis of urban violence in Colombia. Oxf. Bull. Econ. Stat. **39**, 453 (2011)

19. C. Detotto, E. Otranto, Cycles in crime and economy: leading, lagging and coincident behaviors. J. Quant. Criminol. **28**(2), 295–317 (2012)
20. J. Candia, M.C. González, P. Wang, T. Schoenharl, G. Madey, A.-L. Barabási, Uncovering individual and collective human dynamics from mobile phone records. J. Phys. A: Math. Theor. **41**, 224015 (2008)
21. R. Lambiotte, V. Blondel, C. de Kerchove, E. Huens, C. Prieur et al., Geographical dispersal of mobile communication network. Physica A **387**, 5317 (2008)
22. D. Crandall, L. Backstrom, D. Cosley, A. Suri, D. Huttenlocher et al., Inferring social ties from geographic coincidences. Proc. Nat. Acad. Sci. **107**, 22436 (2010)