Disentangling collective trends from local dynamics

Marc Barthélemy[a,b,1], Jean-Pierre Nadal[b,c], and Henri Berestycki[b]

[a]Institut de Physique Théorique (IPhT), Commissariat à l'Energie Atomique (CEA), URA 2306 Centre National de la Recherche Scientifique (CNRS) F11191 Gif-sur-Yvette, France; [b]Centre d'Analyse et de Mathématique Sociales (CAMS), UMR 8557 Centre National de la Recherche Scientifique (CNRS), Ecole des Hautes Etudes en Sciences Sociales (EHESS) 54 Bld Raspail, F-75270 Paris Cedex 06, France; and [c]Laboratoire de Physique Statistique de l'Ecole Normale Supérieure (LPS-ENS), UMR 8550 Centre National de la Recherche Scientifique (CNRS), Université Paris 6 and Paris 7 France

Edited by H. Eugene Stanley, Boston University, Boston, MA, and approved March 11, 2010 (received for review September 8, 2009)

A single social phenomenon (such as crime, unemployment, or birthrate) can be observed through temporal series corresponding to units at different levels (i.e., cities, regions, and countries). Units at a given local level may follow a collective trend imposed by external conditions, but also may display fluctuations of purely local origin. The local behavior is usually computed as the difference between the local data and a global average (e.g, a national average), a viewpoint that can be very misleading. We propose here a method for separating the local dynamics from the global trend in a collection of correlated time series. We take an independent component analysis approach in which we do not assume a small average local contribution in contrast with previously proposed methods. We first test our method on synthetic series generated by correlated random walkers. We then consider crime rate series (in the United States and France) and the evolution of obesity rate in the United States, which are two important examples of societal measures. For the crime rates in the United States, we observe large fluctuations in the transition period of mid-70s during which crime rates increased significantly, whereas since the 80s, the state crime rates are governed by external factors and the importance of local specificities being decreasing. In the case of obesity, our method shows that external factors dominate the evolution of obesity since 2000, and that different states can have different dynamical behavior even if their obesity prevalence is similar.

time series analysis | global trend | crime rate | obesity | independent component analysis

Large complex systems are composed of various interconnected components. The measure of the behavior of a single component thus results from the superimposition of different factors acting at different levels. Common factors such as global trends or external socioeconomic conditions obviously play a role but usually different subunits (such as users of the Internet and states or regions in a country) will react in different ways and add their local dynamics to the collective pattern. For example, the number of downloads on a website depends on factors such as the time of the day but one can also observe fluctuations from a user to another one (1). In the case of criminality, favorable socioeconomic conditions will impose a global decreasing trend whereas local policies will affect the regional time series. In the case of financial series, the market imposes its own trend and some stocks respond to it more or less dramatically. In all these cases it is important to be able to distinguish if the stocks or regions are at the source of their fluctuations or if on the opposite, they just follow the collective trend.

Extracting local effects in a collection of time series is thus a crucial problem in assessing the efficiency of local policies and more generally, for the understanding of the causes of fluctuations. This problem is very general and as the availability of data is always increasing particularly in social sciences, it becomes always more important for the modeling (2) and the understanding of these systems. There is obviously a huge literature on studying stochastic signals (3) ranging from standard methods to more recent ones such as the detrended fluctuation analysis (4), independent component analysis (5–7), and separation of external and internal variables (8, 9). Most of these methods treat the internal dynamics as a small unbiased local perturbation that is in contrast with the method proposed here.

In a first part we present the method. In a second part, we test it on synthetic series generated by correlated random walkers. We then apply the method to empirical data of crime rates in the United States and France, and obesity rates in the United States, for which no general quantitative method is known to provide a separation between global and local trends.

Model and Method

In general, one has a set of time series $\{f_i\}_{i=1,\ldots,N}(t)$ where $t = 1, \ldots, T$. The index i refers to a particular unit on a specific scale such as a region, city, or a country. The problem we address consists of extracting the collective trend and the effect of local contributions. One way to do so is to assume the signal $f_i(t)$ to be of the form

$$f_i(t) = f_i^{\text{ext}}(t) + f_i^{\text{int}}(t) \quad [1]$$

where the "external" part, $f_i^{\text{ext}}(t)$, represents the impact on the region i of a global trend, whereas the "internal" part, f_i^{int}, represents the contribution due to purely local factors. Usually, to discuss the impact of local policies, one compares a regional (local) curve f_i to the average (the national average in case of regions of a country) computed as

$$f^{\text{av}}(t) = (1/N)\sum_i f_i \quad [2]$$

(or $f^{\text{av}} = \sum_i n_i f_i / \sum_i n_i$ if one has intensive variables and populations n_i). Although reasonable at first sight, this assumes that the local component is purely additive: $f_i(t) = f^{\text{av}}(t) + \textit{local term}$. In this article, following (8, 9), we will rather consider the possibility of having both multiplicative and additive contributions. More specifically, we assume

$$f_i^{\text{ext}}(t) = a_i w(t) \quad [3]$$

where $w(t)$ is a collective trend common to all series, and which affects each region i with a corresponding prefactor a_i. These coefficients are assumed to depend weakly on the period considered; i.e., to vary slowly with time. We thus write

$$f_i(t) = a_i w(t) + f_i^{\text{int}}(t). \quad [4]$$

We first note that the global trend w is known up to a multiplicative factor only (one cannot distinguish $a_i w$ from $(a_i z)(w/z)$ whatever $z \neq 0$) and we will come back to this issue of scale later. Also, the purely additive case is recovered if the a_is are independent of i. If on the contrary the a_is are different from one region to the other, the national average **[2]**, $f^{\text{av}} = \bar{f} = (1/N)\sum_i f_i$, is then given by

Author contributions: M.B., J.-P.N., and H.B. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: marc.barthelemy@cea.fr.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.0910259107/-/DCSupplemental.

$$\bar{f}(t) = \bar{a}w(t) + \bar{f}^{\text{int}}. \quad [5]$$

Here and in the following we denote the sample average, that is the average over all units i, by a bar, $\bar{\cdot}$, and the temporal average by brackets $\langle\cdot\rangle$. The "naïve" local contribution is then estimated by the difference with the national average

$$f_i^{\text{int,n}}(t) = f_i(t) - \bar{f}(t) = (a_i - \bar{a})w(t) + f_i^{\text{int}}(t) - \bar{f}^{\text{int}}(t). \quad [6]$$

The estimated local contribution $f_i^{\text{int,n}}(t)$ can thus be very different from the original one, $f_i^{\text{int}}(t)$, and the difference $|f_i^{\text{int,n}}(t) - f_i^{\text{int}}(t)|$ will be very large at all times t where $w(t)$ is large (note that the conclusion would be the same by taking the national average as $f^{\text{av}}(t) = \sum_i n_i f_i / \sum_i n_i$). This demonstrates that comparing local time series with the naive average could in general be very misleading. Beside the correct computation of the external and internal contributions, the existence of both multiplicative and additive local contributions implies that the effect of local policies must be analyzed by considering both how the local unit i follows the global trend (a_i) and how evolves the purely internal contribution (f_i^{int}).

In a previous study (8), de Menezes and Barabasi proposed a simple method to separate the two contributions, internal (f_i^{int}) and external (f_i^{ext} written as $a_i w(t)$). They assume that the temporal average $\langle f_i^{\text{int}}\rangle$ is zero, and compute the external and internal parts by writing

$$a_i = \frac{\sum_t f_i(t)}{\frac{1}{N}\sum_t \sum_j f_j(t)} = \langle f_i\rangle / \langle \bar{f}\rangle \quad [7]$$

and $f_i^{\text{ext}}(t) = a_i \bar{f}(t)$. This method can be shown to be correct in very specific situations, such as the case where f_i is the fluctuating number of random walkers at node i in a network, but in many cases however, one can expect that the local contributions have a non-zero sample average and the method of refs. 8 and 9 will yield incorrect results. Indeed, if the hypothesis [**4**] is exact, this method would give for w the estimate $\widehat{w}(t) = \bar{a}w(t) + \bar{f}^{\text{int}}(t)$, and in the limit $|w(t)| \to \infty$ for $t \to \infty$ would lead to the estimates $\widehat{a}_i \approx a_i/\bar{a}$ and $\widehat{f}_i^{\text{int}} \approx f_i^{\text{int}} - a_i \bar{f}^{\text{int}}/\bar{a}$, which are different from the exact results, except if $\bar{f}^{\text{int}} = 0$.

To separate the two contributions we propose in this article a totally different approach, by taking an independent component analysis point of view in which we do not assume that the local contribution has a zero average (over time and/or over the regions). To express the idea that the internal contribution is by definition what is specifically independent of the global trend, and that the correlations between regions exist essentially only through their dependence in the global trend, we impose that the global trend is statistically independent from local fluctuations

$$\langle w f_i^{\text{int}}\rangle_c = 0 \quad [8]$$

(we denote by $\langle\cdot\rangle_c$ the connected correlation $\langle AB\rangle_c = \langle AB\rangle - \langle A\rangle\langle B\rangle$), and that these local fluctuations are essentially independent from region to region, that is for $i \neq j$

$$\langle f_i^{\text{int}} f_j^{\text{int}}\rangle_c \approx 0 \quad [9]$$

where this statement will be made more precise below. We show that, for large N, these constraints [**8**] and [**9**] are sufficient to extract estimates of the global trend w and of the a_is.

We denote by μ_w the average of w and by σ_w its dispersion, so that we write

$$w(t) = \mu_w + \sigma_w W(t) \quad [10]$$

with $\langle W\rangle = 0$ and $\langle W^2\rangle = 1$. If we denote by $F_i(t) = f_i(t) - \langle f_i\rangle$ and $G_i = f_i^{\text{int}} - \langle f_i^{\text{int}}\rangle$, we have

$$F_i(t) = A_i W(t) + G_i(t) \quad [11]$$

with

$$A_i = a_i \sigma_w. \quad [12]$$

Note that $(\sigma_i^{\text{ext}})^2 \equiv \langle (f_i^{\text{ext}})^2\rangle_c = A_i^2$. If we now consider the correlations between these centered quantities, $C_{ij} = \langle F_i F_j\rangle$, we find

$$C_{ij} = A_i A_j + \langle G_i G_j\rangle. \quad [13]$$

If we assume that for $i \neq j$ $\langle G_i G_j\rangle$ is negligible (of order $1/N$) compared to $A_i A_j$ (which is what we mean by having small correlations between internal components in Eq. **9**), from this last expression we can show that at the dominant order in N, we have

$$\sum_{j/j\neq i} C_{ij} \simeq A_i N \bar{A} \quad [14]$$

$$\sum_{i,j/i\neq j} C_{ij} \simeq N^2 \bar{A}^2. \quad [15]$$

These equations lead to

$$A_i = \frac{\sum_{j/j\neq i} C_{ij}}{(\sum_{jj'/j\neq j'} C_{jj'})^{1/2}} \quad [16]$$

which is valid when $\langle \tilde{G}^2\rangle \ll \bar{A}^2$. We note that our method has a meaning only if strong correlations exist between the different f_is and if it is not the case, the definition of a global trend makes no sense and the approximation used in our calculations are not valid.

In *SI Text* section 1, we show that the factors A_is can also be computed as the components of the eigenvector corresponding to the largest eigenvalue of C_{ij}—a method that is valid under the weaker assumption of having a small number (compared to N) of nondiagonal terms of the matrix $D_{ij} = \langle G_i G_j\rangle$ that are not negligible.

Once the quantities A_i are known, we can compute the global normalized pattern $W(t)$ with the reasonable estimator given by $\overline{F/A}$,

$$W(t) \simeq \frac{\bar{F}}{A}. \quad [17]$$

Indeed,

$$\frac{\bar{F}}{A}(t) = \frac{1}{N}\sum_i \frac{F_i}{A_i} = W(t) + \frac{\bar{G}}{A} \quad [18]$$

and because the quantity $\overline{G/A}$ is a sum of independent variables with zero mean, we can expect it to behave as $1/\sqrt{N}$. We can show that this actually results from the initial assumptions. Indeed, by construction $\langle \overline{G/A}\rangle = 0$ and the second moment is

$$\left\langle \left(\frac{\bar{G}}{A}\right)^2 \right\rangle = \frac{1}{N^2}\sum_{ij} \frac{\langle G_i G_j\rangle}{A_i A_j}. \quad [19]$$

By assumption, we have $\langle G_i G_j\rangle \approx 0$ if $i \neq j$ and we thus obtain $\overline{G/A} \sim 1/\sqrt{N}$.

The computation of the A_is and of W is equivalent to an independent component analysis (ICA) (5–7) with a single source (the global trend) and a large number N of sensors. However, in contrast with the standard ICA, we are not interested in getting only the sources (here the trend W), but also the internal

contributions (which, in a standard ICA framework, would be considered as noise terms, typically assumed to be small). We have already the A_is, and because $W(t)$ has been calculated we can compute $G_i = F_i(t) - A_i W(t)$. We thus obtain at this stage

$$\langle f_i \rangle = A_i \frac{\mu_w}{\sigma_w} + \langle f_i^{\text{int}} \rangle. \quad [20]$$

This is a set of N equations for $N+1$ unknown (μ_w/σ_w and the $\langle f_i^{\text{int}} \rangle$s) and we are thus left with one free parameter, the ratio μ_w/σ_w. Knowing its value would give the N local averages, the $\langle f_i^{\text{int}} \rangle$s. Less importantly, one may want also to fix the average μ_w (hence both μ_w and σ_w) to fully determine the pattern $w(t)$: This will be of interest only for making a direct comparison between this pattern and the national average [**2**]. Eq. **20** suggests a statistical linear correlation between $\langle f_i \rangle$ and A_i, with a slope given by μ_w/σ_w. We will indeed observe a linear correlation in the datasets (see next section). However, it could be that the $\langle f_i^{\text{int}} \rangle$s themselves are correlated with the A_is. Hence, and unfortunately, a linear regression cannot be used to get an unbiased estimate of the parameter μ_w/σ_w. In the absence of additional information or hypothesis this parameter remains arbitrary. However, one may compare the qualitative results obtained for different choices of μ_w/σ_w: which properties are robust, and which ones are fragile. In particular one would like to be able to access how a given region is behaving, compared to another given region, and/or to the global trend. To do so, in the applications below we will in particular analyze: (*i*) the correlations between the two local terms, A_i and $\langle f_i^{\text{int}} \rangle$; (*ii*) the robustness of the rank given by the $\langle f_i^{\text{int}} \rangle$s; (*iii*) the sign of $\langle f_i^{\text{int}} \rangle$; and (*iv*) the quantitative and qualitative similarities between $f_i^{\text{int}}(t)$ and the naive estimate $f_i^{\text{int,n}}(t)$.

We will focus on two particular scenarios. First, one may ask the global trend to fall right in the middle of the N series. There are different ways to quantify this. One way to do so is to note that, in the absence of internal contribution, f_i/a_i would be equal to w, hence $\langle f_i \rangle / A_i$ would be equal to μ_w/σ_w. Therefore we may compute μ_w/σ_w by imposing

$$\frac{\mu_w}{\sigma_w} = \frac{1}{N} \sum_i \frac{\langle f_i \rangle}{A_i}, \quad [21]$$

which is thus equivalent to impose $\frac{1}{N}\sum_i \frac{\langle f_i^{\text{int}} \rangle}{A_i} = 0$. An alternative is to ask the resulting f_i^{int} to be as close as possible to the naive ones [**6**], by minimizing $(1/N)\sum_i \langle (f_i^{\text{int}} - f_i^{\text{int,n}})^2 \rangle$, which gives

$$\frac{\mu_w}{\sigma_w} = \frac{\langle f^{\text{av}} \rangle \bar{A}}{\bar{A}^2}. \quad [22]$$

In both cases one may then fix μ_w from $\mu_w = \langle f^{\text{av}} \rangle$ or by imposing $w(t_0) = f^{\text{av}}(t_0)$ for some arbitrary chosen t_0. Finally, one may rather ask for a conservative comparison with the naive approach by minimizing the difference between w and f^{av}: either by writing $\mu_w = \langle f^{\text{av}} \rangle$ (or $w(t_0) = f^{\text{av}}(t_0)$) and $\sigma_w = \langle (f^{\text{av}})^2 \rangle_c$, or by minimizing $\langle (w - f^{\text{av}})^2 \rangle$, which gives

$$\mu_w = \langle f^{\text{av}} \rangle \quad \text{and} \quad \sigma_w = \langle W f^{\text{av}} \rangle. \quad [23]$$

For N is large, one can check that the results depend weakly on any one of these reasonable choices.

The second scenario considers the correlations between the $\langle f_i^{\text{int}} \rangle$s and the A_is. As we will see, the first hypothesis leads to a strictly negative correlation. An alternative is thus to explore the consequences of assuming no correlations, hence asking for

$$\overline{A \langle f^{\text{int}} \rangle} - \bar{A} \langle \bar{f}^{\text{int}} \rangle = 0, \quad [24]$$

which implies that the slope of the observed linear correlation $\langle f_i \rangle$ with A_i gives the value of μ_w/σ_w. As explained above, for each application below we will discuss the robustness of the results with respect to these choices of the parameter μ_w/σ_w.

We can now summarize our method. It consists of (*i*) estimating the A_is using Eq. **16** [or using the eigenvector corresponding to the largest eigenvalue of the correlation matrix (*SI Text*)], (*ii*) computing W using Eq. **17**, and finally (*iii*) comparing the results for different hypothesis on μ_w/σ_w as discussed above. We propose to call this method the External Trend and Internal Component Analysis (ETICA). We note that if the hypotheses [**4**], [**8**], and [**9**] are correct, the method gives estimates of W, the A_is (hence of $f_i^{\text{int}} - \langle f_i^{\text{int}} \rangle$), which become exact in the limit t and N large, and a good estimate of the full trend w (hence of the $\langle f_i^{\text{int}} \rangle$) whenever this trend, qualitatively, does fall in the middle of the time series.

Once we have extracted with this method the local contribution f_i^{int}, and the collective pattern $w(t)$ together with its redistribution factor a_i for each local series, we can study different quantities, as illustrated below on different applications of the method. In general, although this method gives a pattern $w(t)$ very similar to the sample average $\bar{f}(t)$, we will see that there is nontrivial structure in the prefactors a_is leading to non trivial local contributions $f_i^{\text{int}}(t)$.

In some cases one may expect to have, in addition to the local contribution, a linear combination of several global trends (a small number of "sources"): We leave for future work the extension of our method to several external trends.

Applications: Correlated Random Walkers, Crime Rates in the United States and France, and Obesity in the United States.

We first test our method on synthetic series and we then illustrate it on crime rate series (in the United States and in France) and on US obesity rate series. For the crime rates, a plot of the time series shows that obviously a common trend exists (Fig. 1). After computing the internal and external terms, we perform different tests to assess the validity of the approach. In particular, Fig. 2 shows a plot of the local factors A_is versus the data time-averages, the $\langle f_i \rangle$s. One observes a statistical linear correlation in the four set of time series. We stress that the A_is are computed from the covariance matrix of the data, hence after removing the means from the time series. The fact that we do observe a linear correlation is thus a hint that our hypothesis on the data structure is reasonable (in contrast the very good linear correlation observed in refs. 8 and 9 can be shown to be an artifact of the method used in these works, leading to an exact proportionality independently of the data structure (*SI Text*). We now discuss in more detail the

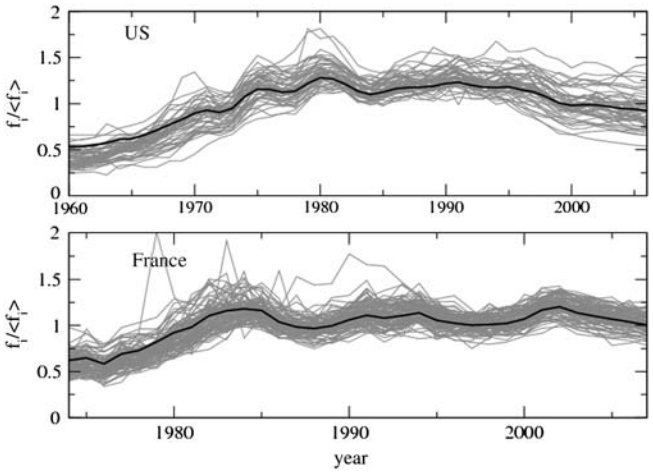

Fig. 1. Collective pattern. Crime rates for the United States (*Upper*) and France (*Lower*) normalized by their time average. The black thick line represent the collective pattern $w(t)$ computed with our method.

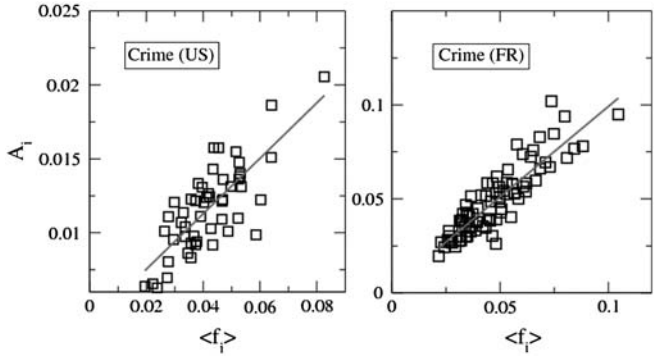

Fig. 2. Existence of a linear correlation. We plot the prefactors A_i versus the time average $\langle f_i \rangle$ for the two crime rate datasets.

synthetic series, each one of the crime rate datasets, and the obesity rate.

Synthetic Series: Correlated Random Walkers. We can illustrate our method on the case of correlated random walkers described by the equation

$$f_i(t) = F(t) + \sum_{\tau=0}^{t} \xi_i(\tau) \qquad [25]$$

where $F(t)$ is the global trend imposed to all walkers and the $\xi_i(t)$ are Gaussian noises of zero mean and with correlations of the form $\overline{\xi_i(t)\xi_j(t)} = [(N-M)\delta_{ij} + \alpha^2 M]/12$ where α and M are tunable parameters (*SI Text*, section 3). For $M = 0$, the random noises $\xi_i(t)$ are independent and our method is very accurate: We choose, for example, a sinusoidal trend $F(t) = \sin(\omega t)$ and we plot, in Fig. 3, the original signal, the exact local contribution, and the local contribution computed with our method. When the correlation between walkers is increasing we study the Pearson correlation coefficient between the original local contribution and the estimate provided by our method, and we observe that our method is indeed accurate as long as the correlations between the G_is are not too large, which corresponds here to the condition $\alpha^2 M \ll 1$ (*SI Text*, section 3).

Crime Rates in the United States and France. In criminology an essential question concerns the impact of local policies, a subject of much debate (10, 11). To assess these local effects (at the level of a state or a region), most authors consider the difference of a state evolution with the national average. As we noticed above, this may lead to incorrect predictions. In this second part of applications, we thus illustrate our method on the analysis of the series of crime rates in 50 states in the United States (12) for the period 1965–2005, and about 100 départements of France (13) for the period 1974–2007. In Fig. 1, we represent these time series normalized by their time average. The observed data collapse confirms the existence of a collective pattern (we also show on this plot the collective pattern $w(t)$ obtained with our method). For the French case, we have withdrawn outliers that do not satisfy our initial assumptions. The series of these départements are indeed uncorrelated with the rest of crime rates and cannot be incorporated in the calculation of the collective pattern. We apply our method to these data and extract $w(t)$, the A_is and $f_i^{\text{int}}(t)$. As already mentioned, we plot, in Fig. 2, the A_is vs. the averages $\langle f_i \rangle$, exhibiting a statistical linear correlation. We can check a posteriori that all conditions assumed in our calculation are fulfilled [zero $\langle w f_i^{\text{int}} \rangle$ and small $\langle G_i G_j \rangle$ (*SI Text*, section 1)]. Also, we checked that the coefficients a_i do not vary too much the period considered, which is an important condition for our method (see discussion on different datasets in *SI Text*, section 4).

To assess quantitatively the importance of local versus external fluctuations, we study in particular the ratio of dispersions defined by

$$\eta_i = \frac{\sigma_i^{\text{ext}}}{\sigma_i^{\text{int}}} \qquad [26]$$

where the external contribution is the standard deviation of $f_i^{\text{ext}}(t) = a_i w(t)$, that is $\sigma_i^{\text{ext}} = A_i$, and the internal one is given by $(\sigma_i^{\text{int}})^2 = \langle (f_i^{\text{int}})^2 \rangle_c = \langle G_i^2 \rangle$. Note that these quantities η_i, being based on fluctuations, does not depend on μ_w/σ_w. This quantity is found in both cases in France and in the United States larger than one. This indicates that external factors always dominate over local fluctuations, whereas local policies seem to play a minor role. In the case of crime, these external effects might be socio-economic factors such as unemployment, density, etc.

In addition, to compute the average of the η_is, we can also observe the time evolution of the heterogeneity defined by the sample variances of the different components. We first observe in Fig. 4 that large fluctuations are observed in the transition period of mid-70s during which crime rates increased significantly.

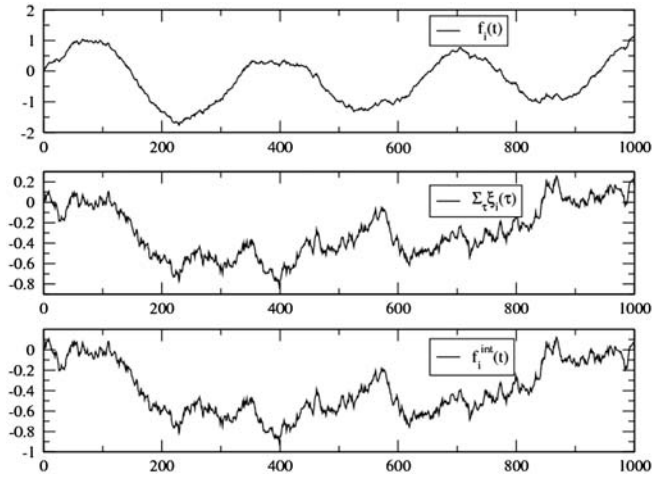

Fig. 3. (*A*) Original signal composed of the superimposition of a sinusoidal trend and Gaussian noises (for $N = 100$ walkers). (*B*) Exact local contribution. (*C*) Local contribution extracted with our method.

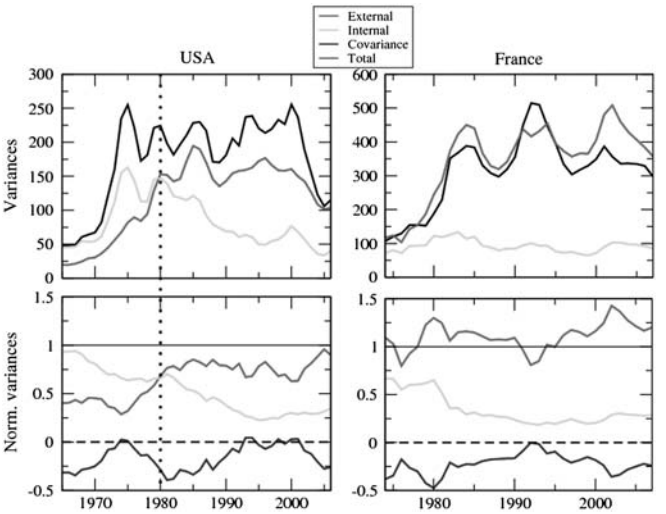

Fig. 4. Comparison of internal and external fluctuations. We present results for the United States (*Left Column*) and France (*Right Column*). In the upper panels, we represent the total variance of the signal, the external and the internal contribution. In the lower panels, we represent the external, internal, and the covariances normalized by the variance of the signal. We can observe that for the United States, the external contribution is dominating since the 80s.

We also observe for the United States that until 1980, fluctuations were essentially governed by local effects but that this trend is inverted and increases in the period post-1980s. In particular, during the period 1980–2000, during which one observes a decline of crime rates (11), it is the collective trend which determines the fluctuations.

Even if we have presented results for reasonable choices of the parameter σ_w (in the following we make the harmless choice $\mu_w = 1$), one can ask the question of the robustness of different observed properties. First, we can compare the predictions for σ_w obtained for the different assumptions used in this paper. In the upper panels for Figs. 5 and 6 we show for the United States (France), the quantities $\langle f_i^{\rm int}\rangle/a_i$, $\langle f_i^{\rm int}\rangle/\overline{a}$, and $r = (\langle f_i^{\rm int}\rangle a - \langle f_i^{\rm int}\rangle\,\overline{a})/\sigma_a^2$.

We see in these figures that these quantities are zero for values of σ_w, which are very close. We also compute the fraction of time p_i for which $f_i^{\rm int}(t)$ and the naive calculation $\langle f_i\rangle - \langle f_i^{\rm av}\rangle$ have different signs. We plot in the lower panels of Figs. 4 and 5, the quantity $p = \frac{1}{N}\sum_i p_i$ showing that for this range of σ_w, the signs of $\langle f_i^{\rm int}\rangle$ and $\langle f_i\rangle - \langle f_i^{\rm av}\rangle$ are the same for about 60% of the time period. We can also study the sign $\langle f_i^{\rm int}\rangle$ vs. σ_w and we can observe some robustness. In particular, in the US case, approximately 6 states (CA, NV, MO, MI, NY, and AZ) have a positive local contribution (in the range $\sigma_w \in [0.24, 0.32]$ whereas 6 states have always a negative local contribution (VT, GA, LA, NH, CT, and MS). In these cases we can reasonably imagine that local policies have a noticeable effect.

Finally, we can also analyze the ranking of the local contributions $\langle f_i^{\rm int}\rangle$ vs. σ_w by studying Kendall's τ for the two consecutive series $\{\langle f_i^{\rm int}\rangle\}(\sigma_w)$ and $\{\langle f_i^{\rm int}\rangle\}(\sigma_w + \delta\sigma_w)$. In both cases (France and United States) we observe a value $\tau > 0.9$ for the range chosen $\sigma_w \in [0, 0.5]$ (the control case for a random permutation being <0.1) indicating a large robustness of the ranking. This means that independently of the assumption used to compute σ_w we can rank the different regions according to the importance of their local contribution.

Obesity in the United States. The prevalence of obesity (defined as a body mass index—BMI, which is the ratio of the body mass to the square of the height—larger than 30 kg/m^2) is rapidly increasing in the world (14) and reached epidemic proportion in the United States and is now a major public health concern (15–17).

Disparities by sex and between ethnic groups have been observed in the prevalence of obesity (18), but few studies focus on the effect of local factors and policies on the obesity rate.

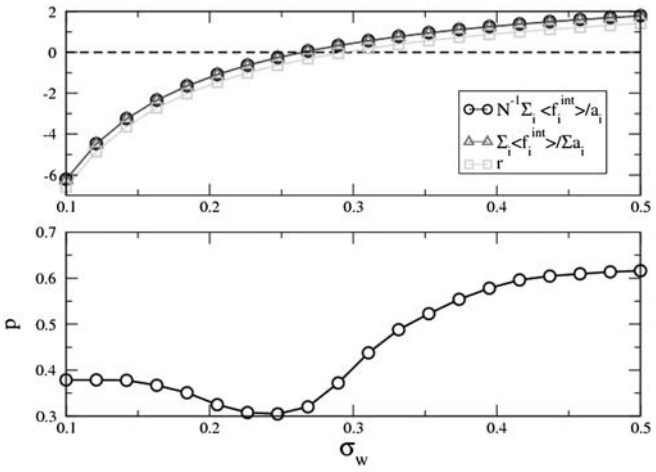

Fig. 5. Determination of σ_w in the US crime rate case. We can use various conditions to determine σ_w: $0 = N^{-1}\sum_i \langle f_i^{\rm int}\rangle/a_i$, $0 = \sum_i \langle f_i^{\rm int}\rangle/\sum_i a_i$, or $r = 0$ (r is defined in the text). We see in this plot that they all give very similar values. (*Lower Panels*) Average fraction of time for which $\langle f_i^{\rm int}\rangle$ has the same sign as the naive calculation $\langle f_i\rangle - \langle f_i^{\rm av}\rangle$.

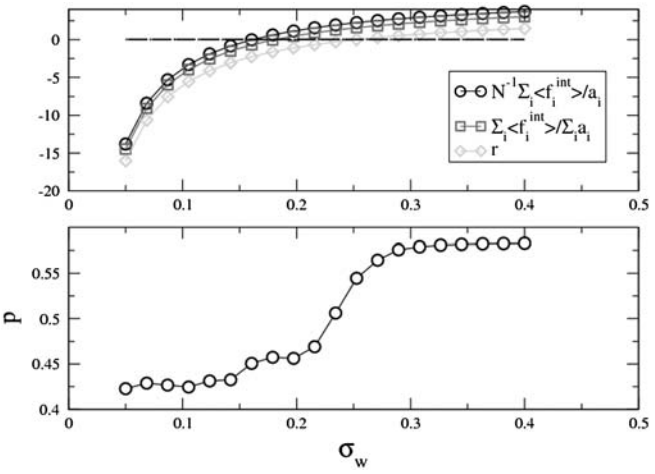

Fig. 6. As in Fig. 5, we can determine σ_w in the case of the crime rate in France, by using different conditions: $0 = N^{-1}\sum_i \langle f_i^{\rm int}\rangle/a_i$, or $0 = \sum_i \langle f_i^{\rm int}\rangle/\sum_i a_i$, or $r = 0$. Here also, these conditions give very similar values of σ_w. (*Lower Panels*) Average fraction of time for which $\langle f_i^{\rm int}\rangle$ has the same sign as the naive calculation $\langle f_i\rangle - \langle f_i^{\rm av}\rangle$.

We thus apply our method to data from the Center for Disease Control and Prevention (19), which describe the percentage of the population that is obese for each state in the United States and for the period 1995–2008. As in the crime rate case, we can compare the variances for the internal and external contributions (*SI Text*, section 5) and we observe that the external contribution is dominating since the year 2000. This result means that the global trend is the major cause of the evolution of obesity in different states. We can get more detailed information about the specific behavior of the states by studying the ratio η_i defined in Eq. **26** and the ratio y_i of the time averages of the local contribution and the total signal $y_i = \langle f_i^{\rm int}\rangle/\langle f_i\rangle$. We represent these two quantities in a plane (Fig. 7) and we first note that for all states $\eta_i > 1$, which means that fluctuations are mainly governed by the global trend. We can also divide the states into two groups with $y_i > 0$ and $y_i < 0$. For large and positive y_i (such as DC and IN for example), the states have a small a_i, which means that these states are the less susceptible to the global trend, whereas in the opposite case (such as GA or AZ), the states are governed by the global trend. Within each group we can then distinguish the states according to their level of fluctuations (η_i close to or much larger than one).

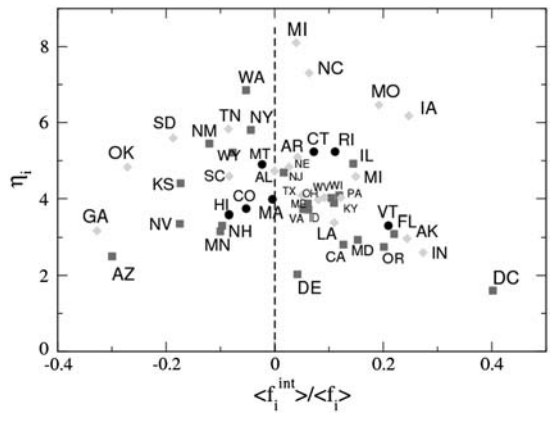

Fig. 7. Fluctuations versus importance of the local contribution. We plot the quantity η_i versus $y_i = \langle f_i^{\rm int}\rangle/\langle f_i\rangle$ for the United States. We divide the states in three groups: obese percentage <22% (*Circles*), percentage in the interval 22 and 26% (*Squares*), and percentage >26% (*Diamonds*). Low prevalence states seem to concentrate in the same region $y_i \approx 0$, whereas medium- and large-prevalence states display very different values of η_i and y_i.

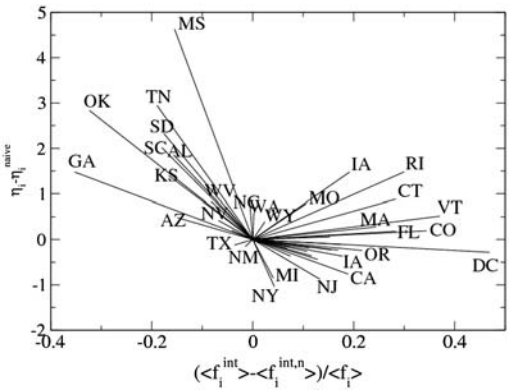

Fig. 8. Difference with the naive fluctuations and local contribution. We represent for the different states the difference vectors $(\langle f_i^{\text{int}} - f_i^{\text{int,n}}\rangle / \langle f_i\rangle, \eta_i - \eta_i^{\text{naive}})$ where $\eta_i^{\text{naive}} = \langle \langle (f^{\text{av}})^2\rangle_c / \langle (f_i^{\text{int,n}})^2\rangle_c$. We observe that for about half of the states the difference between the naive calculation and our method is not negligible {for the sake of clarity, we indicated the name of the corresponding state for most vectors except for those close to (0,0)].

The states Arizona, Georgia, and Oklahoma for example have all a local contribution of the same order but their fluctuations properties are different (with larger external fluctuations for OK, for example). More generally, we can see in this figure that states with large prevalence display very different values of (y_i, η_i). This result points toward the fact that describing states by their prevalence only can be very misleading and can hide important dynamical behaviors. Finally, we also computed the quantities y_i and η_i using the naive local contribution defined by $f_i^{\text{int,n}}(t) = f_i(t) - f^{\text{av}}(t)$ (see Eqs. **2** and **6**). We represent, in Fig. 8, the difference as vectors of components given by $(\langle f_i^{\text{int}} - f_i^{\text{int,n}}\rangle / \langle f_i\rangle, \eta_i - \eta_i^{\text{naive}})$ and we can see in this figure that for roughly half of the states the naive calculation of the local contribution can be very misleading.

Discussion

In this article we adressed the crucial problem of extracting the local components of a system governed by a global trend. In this case, comparing the local signal to the average is very misleading and can lead to wrong conclusions. We applied this method to the example of crime rates series in the United States and France and our analysis revealed surprising facts. The important result is about the importance of fluctuations, which after the 80s in the United States are governed by external factors. This result suggest that understanding the evolution of crime rates relies mostly on the identification of global socioeconomic behavior and not on local effects such as state policies etc. In particular, this result could also help in understanding the decreasing trend observed in the United States and which so far remains a puzzle (11, 20). In the case of obesity, we show that since the year 2000, external factors dominate, and maybe more importantly that states with the same level of prevalence have very different dynamical behaviors, thus calling for the need of a detailed study state by state.

However, one may expect an even better signal analysis by assuming that there are several independent external trends: It will be interesting to see if our approach, combined with the more standard ICA techniques, can be generalized to the case of several global trends (a small number of sources). The recent availability of large amounts of data in social systems call for the need of tools able to analyze them and to extract meaningful information and we hope that our present contribution will help in the understanding of these systems where the local dynamics is superimposed to collective trends.

ACKNOWLEDGMENTS. We thank the anonymous referees for constructive remarks, in particular about the applicability conditions of our method. This work is part of the project "DyXi" supported by the French National Research Agency Grant ANR-08-SYSC-008.

1. Huberman BA (2001) *The Laws of the Web* (MIT Press, Cambridge, MA).
2. Castellano C, Fortunato S, Loreto V (2009) Statistical physics of social dynamics. *Rev Mod Phys* 81:591–646.
3. Kautz H, Schreiber T (1997) *Nonlinear time series analysis* (Cambridge University Press, Cambridge, U.K.).
4. Peng CK, et al. (1994) Mosaic organization of DNA nucleotides. *Phys Rev E* 49:1685–1689.
5. Comon P (1994) Independent component analysis: A new concept? *Signal Process* 36(3):287–314.
6. Cardoso J-F (1997) Statistical principles of source separation. *Proceedings of the 11th IFAC Symposium on System Identification* (IFAC, Fukuoka, Japan), pp 1837–1844.
7. Hyvrinen AJ, Karhunen J, Oja E (2001) *Independent Component Analysis* (Wiley, New York).
8. de Menezes MA, Barabasi A-L (2004) Fluctuations in network dynamics. *Phys Rev Lett* 92:028701.
9. de Menezes MA, Barabasi A-L (2004) Separating internal and external dynamics of complex systems. *Phys Rev Lett* 93:068701.
10. de Maillard J, Roché S (2004) Crime and justice in France: Time trends, policies and political debate. *European Journal of Criminology* 1:111–151.
11. Zimring F (2007) *The Great American Crime Decline* (Oxford University Press, Oxford).
12. United States Uniform Crime Report—State Statistics from 1960–2007. Available at http://www.disastercenter.com/crime/.
13. France: Institut National des Hautes Etudes de Sécurité La Documentation française: Criminalité et délinquance constatées en France—Tome I : données générales, nationales, régionales et départementales. Available at http://www.inhes.interieur.gouv.fr/Bulletin-annuel-112.html and http://www.ladocumentationfrancaise.fr/rapports-publics/084000201/.
14. James PT, Leach R, Kalamara E, Shayeghi M (2001) The worldwide obesity epidemic. *Obes Res* 9:228S–233S.
15. Mokdad AH, et al. (1999) The spread of obesity epidemic in the United States, 1991–1998. *J Amer Med Assoc* 282:1519–1522.
16. Ogden CL, et al. (2006) Prevalence of overweight and obesity in the United States 1999–2004. *J Amer Med Assoc* 295:1549–1555.
17. Hedley AA (2004) Prevalence of overweight and obesity among US children, adolescents, and adults 1999–2002. *J Amer Med Assoc* 291:2847–2850.
18. Wang Y, Beydoun MA (2007) The obesity epidemic in the United States—gender, age, socioeconomic, racial/ethnic, and geographic characteristics: A systematic review and meta-regression analysis. *Epidemiol Rev* 29:6–28.
19. Centers for Disease Control and Prevention Behavioral Risk Factor Surveillance System. Available at http://apps.nccd.cdc.gov/brfss/.
20. Levitt SD (2004) Understanding why crime fell in the 1990s: Four factors that explain the decline and six that do not. *J Econ Perspect* 18:163–190.