

# Computational Socioeconomics

Jian Gao<sup>a,b,c</sup>, Yi-Cheng Zhang<sup>d,e,\*</sup>, Tao Zhou<sup>a,c,f,\*\*</sup>

<sup>a</sup>*CompleX Lab, University of Electronic Science and Technology of China, Chengdu 611731, PR China.*

<sup>b</sup>*MIT Media Lab, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.*

<sup>c</sup>*Big Data Research Center, University of Electronic Science and Technology of China, Chengdu 611731, PR China.*

<sup>d</sup>*Institute of Fundamental and Frontier Science, University of Electronic Science and Technology of China, Chengdu 610054, PR China.*

<sup>e</sup>*Department of Physics, University of Fribourg, CH-1700 Fribourg, Switzerland.*

<sup>f</sup>*Institution of New Economic Development, Chengdu 610110, PR China.*

## Abstract

Uncovering the structure of socioeconomic systems and timely estimation of socioeconomic status are significant for economic development. The understanding of socioeconomic processes provides foundations to quantify global economic development, to map regional industrial structure, and to infer individual socioeconomic status. In this review, we will make a brief manifesto about a new interdisciplinary research field named *Computational Socioeconomics*, followed by detailed introduction about data resources, computational tools, data-driven methods, theoretical models and novel applications at multiple resolutions, including the quantification of global economic inequality and complexity, the map of regional industrial structure and urban perception, the estimation of individual socioeconomic status and demographic, and the real-time monitoring of emergent events. This review, together with pioneering works we have highlighted, will draw increasing interdisciplinary attentions and induce a methodological shift in future socioeconomic studies.

**Keywords:** Socioeconomics; network science; data mining; machine learning

## Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>Global development, inequality and complexity</b>	<b>6</b>
2.1	World development and poverty mapping . . . . .	6
2.1.1	Remote sensing observes poverty . . . . .	6
2.1.2	Mobile phones reveal socioeconomic status . . . . .	9
2.1.3	Combined data for better inference . . . . .	10
2.2	Economic complexity and fitness of nations . . . . .	12
2.2.1	Product space and economic complexity . . . . .	12
2.2.2	Fitness index and economic dynamics . . . . .	14
2.2.3	Variant indices and development analysis . . . . .	16
2.3	Spatial demography and culture evolution . . . . .	19
2.3.1	World population distribution . . . . .	19
2.3.2	International migration . . . . .	21
2.3.3	Culture evolution . . . . .	23

\*Corresponding author at: Department of Physics, University of Fribourg, CH-1700 Fribourg, Switzerland.

\*\*Corresponding author at: CompleX Lab, University of Electronic Science and Technology of China, Chengdu 611731, PR China.

Email addresses: gaojian08@hotmail.com (Jian Gao), yi-cheng.zhang@unifr.ch (Yi-Cheng Zhang), zhutou@ustc.edu (Tao Zhou)

<b>3 Regional socioeconomic status and urban perception</b>	<b>27</b>
3.1 Economic activity and socioeconomic status . . . . .	27
3.1.1 Nighttime lights reflect economic activity . . . . .	27
3.1.2 Very high resolution imagery maps poverty . . . . .	28
3.1.3 Mobile phones track socioeconomic levels . . . . .	30
3.1.4 Social media reveals socioeconomic status . . . . .	32
3.2 Industrial structure and development path . . . . .	34
3.2.1 Economic structure and relatedness . . . . .	34
3.2.2 Collective learning in economic development . . . . .	36
3.2.3 Development paths and strategies . . . . .	39
3.3 Urban scalings and perception . . . . .	41
3.3.1 Scaling laws for cities . . . . .	41
3.3.2 Unfolding urban functional areas . . . . .	45
3.3.3 Perceiving urban environment . . . . .	48
3.3.4 Urban computing for better lives . . . . .	51
<b>4 Individual socioeconomic status and attributes</b>	<b>54</b>
4.1 Individual socioeconomic level . . . . .	54
4.1.1 Mobile phone and credit card usage . . . . .	54
4.1.2 Social profile and network structure . . . . .	56
4.1.3 Human mobility pattern . . . . .	59
4.2 Employment and performance . . . . .	62
4.2.1 Search queries indicate unemployment . . . . .	62
4.2.2 Other sources relevant to unemployment . . . . .	64
4.2.3 Individual and group performance . . . . .	66
4.3 Demographics and personal variables . . . . .	69
4.3.1 Demographic inference . . . . .	69
4.3.2 Personality analysis . . . . .	71
4.3.3 Online reputation evaluation . . . . .	73
4.3.4 Emotion and health analysis . . . . .	76
<b>5 Situational awareness and disaster management</b>	<b>79</b>
5.1 Public health and epidemic surveillance . . . . .	79
5.1.1 Search queries for epidemic surveillance . . . . .	79
5.1.2 Online posts for disease surveillance . . . . .	81
5.1.3 Mobile phone records for epidemic prediction . . . . .	83
5.2 Emergency and disaster monitoring . . . . .	86
5.2.1 Remote sensing for disaster assessment . . . . .	86
5.2.2 Mobile phones for emergency management . . . . .	88
5.2.3 Social media for situational awareness . . . . .	91
<b>6 Discussions</b>	<b>94</b>
<b>Acknowledgements</b>	<b>96</b>
<b>References</b>	<b>96</b>

## 1. Introduction

Many branches of science have experienced the paradigm shift from qualitative to quantitative studies. Even the most representative one for quantitative sciences, physical science, has undergone a long period for qualitative explorations in its early stages. For example, more than two thousand years ago, Aristotle raised the famous *four elements theory*, which claims that the four classical elements, namely earth, water, air and fire, are the material basis of the physical world. At almost the same time, some Chinese ancient philosophers proposed the *Wu Xing theory* (i.e., the Chinese five elements theory), which is a fivefold conceptual scheme that uses the proportion of ingredients and movements of the five elements (i.e., metal, wood, water, fire and earth) to explain a wide array of phenomena, from the cosmic cycles to the validity of a dynasty. For about two thousand years, the ancient Greek system, contributed by Aristotle and some others, represents the most advanced understanding of the world, which is indeed one of the most influential theory in human history. Up to the end of the middle ages, thanks to the quantitative analyses and experimental verifications, these ancient theories, such as Aristotle's four elements theory and kinetic theory, were progressively replaced by modern scientific theories like the Atomic theory and the Newton's laws.

In contrast to physical science that concentrates on the study of matter and its motion through space and time, social science investigates the social structure based on the activities of and relations between human beings, including sociology, economics, politics, linguistics, jurisprudence, and many other branches. In comparison with physical science, the way from qualitative to quantitative studies is more difficult for social science. On the one hand, the objects under social science study are much more complex than those under physical science study. An individual person is one of the most important units for social science study, playing an analogous role to an atom in physical science [1]. However, human behaviors exhibit heterogeneity and burstiness: different people have much different behavioral patterns and even the same person shows far different behaviors in different spaces and times [2]. Therefore, except a certain success in analyzing the flow of human crowds [3, 4], to treat human beings as atoms will kill many interesting social phenomena. Some other objects under study are naturally not easy to be characterized numerically, such as policies and legal provisions. On the other hand, social science study inevitably suffers from uncertainty and incompleteness. The factors affecting social development are countless, and thus any seemingly coverall theory cannot include all relevant factors and be self-contained. In addition, every single factor is unstable and not independent, being affected by other factors and the external environment. The above intrinsic complexity makes it infeasible to quantitatively test and verify any social theory through controllable repeated experiments in a closed environment, while such experimental verification is indeed the methodological cornerstone that pushes forward physical science and other branches of natural science [5]. At the same time, social science is not good at quantitative predictions to the future, with many predictions from experts and complicated theories being no better than wild guesses [6]. What a pity is that such incorrect predictions cannot subvert the corresponding social theories (much different from physical science) since the mistakes are attributed to the unknown/undetected factors or emergent events [7], instead of the flaws of the theories themselves.

Up to now, along with the development of quantitative methods, social science has successfully learned how to be wise after the event. That is to say, we can always find some theoretical models (possibly together with some cosmetic changes) to provide qualitatively correct or even quantitatively accurate explanations after the event. However, these theories are usually powerless in predicting the future. Confronting such straits, social scientists should not turn back to the qualitative description, but insist on quantitative explanation and prediction, and evaluate the validity of a theory based on its explanatory power and prediction accuracy before the event. In fact, social science study recently shows higher and higher level of quantification and becomes increasingly dependent on real data [8, 9]. However, the traditional way to obtain real data has many limitations. For example, survey data from questionnaires and self-reports usually contains a small number of samples and suffers from social desirability bias (i.e., subjects tend to give socially acceptable answers, instead of the real facts) [10]. Larger-scale and more precise data, such as data from economic census, usually consumes huge resources and lacks timeliness. In many poor countries and regions, population-scale economic census is not feasible. Fortunately, thanks to the digital wave that sweeps across the whole world [11], social scientists have an unprecedented opportunity to develop a quantitative methodology. Indeed, it is for the first time in history, data in the processes of social and economic development, as well as the data of human activities, are recorded by more and more sensing devices, online platforms and other data acquisition terminals. However, these data are not well-structured and are different from the normally handled data in social science. Typical examples include satellite remote sensing data, mobile phone data, social media data, and so on. On the one hand, to understand and analyze

these data asks for advanced techniques in data mining and machine learning, which is a considerable challenge to traditional social scientists. On the other hand, these data are of larger size, almost in real time and with higher resolution, which can reduce the sparsity and bias in small-size data, and reduce the invisible parts in the developing processes (e.g., data points in two consecutive censuses are usually across a few years, and the changes in between are not visible). Therefore, based on these large-scale novel data, we can in principle make great progress in perceiving socioeconomic situations, evaluating and amending known theories, enlightening and creating new theories, detecting abnormal events, predicting future trends, and so on.

The above-mentioned challenges and corresponding attempts have led to the emergence of a new scientific branch, which studies various phenomena in socioeconomic development by using quantitative methods that based on large-scale real data, with particular attention to the economic development problems related to social processes and the social problems related to economic development. We name it as *Computational Socioeconomics*, which is immature, but future-pointing and burgeoning. The computational socioeconomics can be considered as a new branch of socioeconomics resulted from the transformation of methodology, or as a new branch of computational social science by emphasizing on socioeconomic problems.

In the above definition, three keywords are worth paying close attention to. The first one is “quantitative methods”, which emphasizes the usage of numerical values, rather than qualitative description, in characterizing problems and presenting results. In the 5th century BC, the ancient Greek doctor Hippocrates (who is often referred to as the “Father of Medicine”) proposed the *four temperaments theory*, which suggests that there are four fundamental personality types: sanguine, choleric, melancholic, and phlegmatic, and the personality type of an individual is determined by the excess or lack of four body fluids: blood, yellow bile, black bile, and phlegm [12]. Such a qualitative theory, analogous to the impacts of the four elements theory on physical science, has ruled social psychology (in particular the studies on personality) for more than two thousand years. In despite of some reasonable ingredients, the four temperaments theory has stayed on the level of qualitative description, and thus failed to accumulate scientifically solid achievements in its long-time development. Only after modern psychologists obtained quantitative evaluations of the Big Five personality traits via standard scales, personality analysis became an important research domain that plays central roles in many issues of social psychology [13]. Such example show the importance and necessity of the development of quantitative methods. The second one is “real data”, which emphasizes that any theoretical model should respect real data and use the explanatory power and prediction accuracy for real data as the evaluation criteria for its validity. Economics shows a high level of quantification, with most theoretical models being precisely described by a group of elegant equations. Accordingly, given the values of necessary parameters, many targeted economic variables are calculable. However, the majority of economic theories have cocooned themselves in a quantitative fantasyland consisted of ideal assumptions while largely ignored real data. It eventually makes the classical economic theories beautiful rather than practical. For the short term, it cannot predict the upcoming economic crisis [14] (but it can always find out graceful and reasonable theoretical explanations after the crisis [15]). For a long time, it failed to provide effective strategies on economic development for more than a hundred of developing countries over the world [16]. The third one is “large-scale”, which emphasizes the importance of population-scale data (i.e., the data that can directly reflect the entire population under study, instead of a small sample). A very small data set may not only bring statistical bias, but result in completely wrong conclusions. For example, a widely accepted theory by academic community, which has also been validated by various experiments on small-scale social networks, is that the interacting strength between two connected individuals (which can be measured by the frequency and duration of mobile communication or the number of comments, replies and mentions on a social platform, and so on) decays as the increase of the range of their link (the range of a link is defined as the shortest distance between its two endpoints after the removal of this link, and a large link range indicates that the two corresponding endpoints locate in two distant communities with few overlapping nodes) [17, 18]. A very recent experiment on 11 population-scale social networks, however, shows that the interacting strengths through very long-range links are not weaker than those through short-range links [19], which fundamentally challenges our traditional understanding of social network organization.

In comparison with routine methods in social science, the increasing diversity and volume of data lead to methodological changes in two aspects. Firstly, simple statistical tools are not suitable for analyzing unstructured data, such as remote sensing images, street views, social networks, textual content, and so on. Therefore, researchers are badly in need of artificial intelligence, in particular advanced techniques of data mining and machine learning, such as deep learning [20]. Secondly, with population-scale data, sampling is not a necessary method to estimate the statistical properties of the whole population. Instead, one can concentrate on a small-size subset sampled from the original

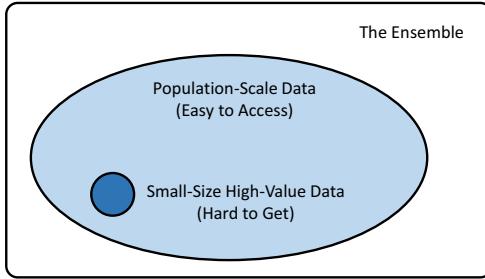


Figure 1: Illustration of the relationship between the entire population, the easily accessible population-scale data and the small-size high-value data. The small-size data set contains some high-value data dimensions that do not appear in the original population-scale data.

data in hand, and add new dimensions of data. These new data dimensions are usually of high values, which can be obtained from traditional ways like manual labelling and questionnaire survey. Using such a small sample as the training data, one can learn a model to infer new dimensions of data from the original dimensions. Applying such model to the whole data set, new dimensions of data for all individuals appeared in the original dataset can be obtained in principle. Such method integrates some routine methods like sampling, labelling and surveying, while it is much more powerful. For example, it is relatively easy to obtain the population-scale data on mobile communication and mobility (all can be obtained from mobile phones), in contrast, it is very hard to know the household income of every family since a poor country cannot support a population-scale economic census and such data is usually treated as official secrets that are not open for public or research institutions. Under the circumstances, we can obtain household incomes of a certain number of families (what we need is just a tiny fraction of all families) via routine questionnaires. These much smaller data set can be used as training data, based on which we can apply machine learning techniques to build a model that can predict household income of a family from the mobile phone data of the family members [21]. Although the inferred data is not perfect, it can be very close to the real data under a certain well-designed algorithm. Notice that, a significant advantage is that the high-value data for almost every individual can be obtained at a very low cost. As shown in Figure 1, combining the accessible population-scale data, a small sample of high-value but hard-to-get data, and a properly selected or well-designed algorithm to infer the high-value data for individuals other than the sample is a novel and representative method in the computational socioeconomics study, showing the deep integration of social science and computer science methods.

Long-term speaking, no matter computational socioeconomics will become a mature branch of science with distinct borderlines or it will completely integrate into the framework of traditional social science, the above-mentioned novel perspective and methodology, driven by big data and artificial intelligence, will definitely become the mainstream in the future and change the landscape of science research in a profound and irrevocable way. Inspired by this positive judgment, we decide to present this review article. In addition, there are three technical reasons for us to write this review. Firstly, computational socioeconomics is an emerging research domain with research findings published in disparate journals and conference proceedings across many disciplines. Therefore, it is necessary to collect these results together. Secondly, we would like to sort and classify representative results according to the objects and data sets under study, so that it is easy for readers to see the landscapes of both methods and achievements. A proper taxonomy can largely reduce the difficulty to master the related knowledge and methods. Although the presented one is built just according to the current progresses of this field, it will evolve to be a more systematic and reasonable one along with further studies. Thirdly, in the nascent stage of computational socioeconomics, different research articles used different expressions to describe essentially the same problems and methods, and thus it is valuable to unify the problem description and the symbolic system. In a word, we hope this review will become a handbook for researchers who are willing to contribute to the development of computational socioeconomics. Furthermore, the paradigm shift in methodology, as presented in this review, is not only relevant to socioeconomics, but also to most branches of social science and to many other qualitative disciplines beyond social science.

The remainder of this review article is organized as follows. The second section will discuss some important problems at the macroscopic scale, such as the world economic development, the competitive powers of countries, the inequality problem, and so on. The third section will mainly concentrate on the urban scale and introduce some

novel ways to solve problems related to the regional economic development, such as how to precisely perceive regional socioeconomic status and how to choose the suitable development paths and strategies for a city. The fourth section will focus on individual level, discussing how to make use of some unobtrusive data to estimate the individual socioeconomic status, including income, employment situation, and even health condition. The fifth section will go a little beyond the scope of computational socioeconomics, and to discuss how the frequently-used data in the previous sections can be utilized to benefit the emergency management and disaster assistance. We cover such issue because the emergency management is an increasingly important social problem and the reported methods are consistent to the methods introduced in the previous sections. In this review, many different data resources have played important roles, among which the following three are the most important: remote sensing satellites, mobile phones and social media platforms. For each of the three data resources, there are some certain representative analytics tools and methods, so it is very effective to sort the results according to the data resources and corresponding methods. Finally, in the last section, we will summarize representative progresses, explore the tendency of the development of computational socioeconomics, discuss the challenges and opportunities in this emerging field, and outline some potentially interesting and significant open issues.

## 2. Global development, inequality and complexity

### 2.1. World development and poverty mapping

Revealing the status of economic development is one of the long-standing problems in socioeconomics [22, 23]. Recently, data with improved quantity and quality have been used to map nations' economic characteristics such as poverty, which comes with economic development and is a major cause of societal instability. Based on an international poverty line at USD 1.25 a day in 2008 [24], 1.2 billion people (21%) lived in poverty in 2012 [25]. Reducing poverty is thus a key target of the Millennium Development Goals (MDGs). To approach this goal, the first step is to accurately map the spatial distribution of poverty [26]. New data and tools have been utilized to better reveal, explain and predict global poverty and economic inequality. In this section, we will briefly introduce literature that map poverty from satellite imagery, infer socioeconomic status from mobile phone (MP) data and fight against poverty with combined data.

#### 2.1.1. Remote sensing observes poverty

Remote sensing (RS) is the acquisition of information by using sensor technologies to detect objects on earth, which is originally used in earth science disciplines [27]. In recent years, high resolution data from RS, for example, nighttime lights (NTLs) satellite imagery, has been used to supply information about economic activity, especially in developing countries where traditional economic census data are insufficient [28]. With a great potential for recording the presence of humanity on the surface of earth, NTLs data can provide an unambiguous indication of the spatial distribution of economic development. Indeed, NTLs have been found to be a powerful predictor of ambient population density and economic activity. Nightsat [29] is a concept for a satellite system, which is capable of global observation. The Nightsat can capture the location and density of lighted infrastructures within human settlement.

One of the pioneering works by Elvidge et al. [30] suggested that NTLs data can be used as a proxy for socioeconomic development in developing countries. Lighted area has a high correlation with the gross domestic product (GDP) and electric power consumption. Moreover, lighted area is strongly correlated with GDP for 21 countries. Latter, by combining lighted area with ancillary statistical information of a city, Doll et al. [31] investigated the potential of NTLs data for quantitative estimation of global socioeconomic parameters. They found that the country-level total lighted area exhibits significantly high correlations with GDP and CO<sub>2</sub> emission. Sutton and Costanza [32] estimated the amount of light energy (LE) from satellite images with global coverage at a high spatial resolution. They found that LE is correlated with GDP at the national level and can serve as a more accurate indicator of economic activity. That is because LE is more spatially explicit and can be directly observed and easily updated almost in real time.

Together with census and survey data, NTLs data have been applied in mapping poverty. Ebener et al. [33] applied regression methods using NTLs imagery to model the distribution of wealth within 171 countries at the national level and 26 countries at the subnational level. They showed that NTLs data is correlated with GDP per capita and other socioeconomic indicators. Noor et al. [34] computed asset-based poverty by applying the principal component analysis (PCA) to NTLs data and household survey data of 37 countries in Africa. They found that the mean brightness

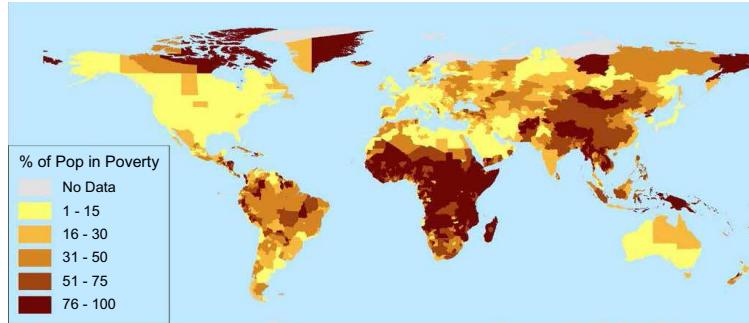


Figure 2: Percentage of population in poverty for subnational administrative units. The world poverty map was estimated based on satellite data-derived poverty index. Figure from [36].

of NTLs data can offer a robust and inexpensive alternative to asset-based poverty indices derived from survey data, suggesting that it is possible to explore and track economic inequity at subnational levels by leveraging NTLs data. For Uganda, Rogers et al. [35] presented a discriminant analysis model to predict poverty after combining satellite imagery and household survey data. They estimated the poverty index by the likelihood of each pixel falling within a specified poverty class. They found that external and independent data have descriptive power for poverty mapping. These novel data sources are likely to outperform socioeconomic datasets that are internally correlated and exploited by small area methods.

A spatially disaggregated global poverty map of 233 countries was produced by Elvidge et al. [36]. The poverty levels were estimated by dividing the LandScan population count data [37] by the brightness from NTLs data (see Figure 2). The produced poverty indices correlate very strongly with other widely accepted measures, suggesting that satellite imagery can enhance the knowledge of socioeconomic conditions around the world at a fine spatial resolution. Later, Ghosh et al. [38] proposed a model to estimate the world-wide economic activity. In their model, a grid of nonagricultural economic activity was created according to the NTLs, while a grid of agricultural activity was created according to the LandScan population grid. Then, by integrating the two grids, a disaggregated map of total economic activity was produced, which can provide an alternative means for measuring global economic activity and predicting future socioeconomic trends.

To better estimate true income growth from NTLs, Henderson et al. [39] developed a statistical framework to estimate two parameters. One is a coefficient that maps NTLs growth into a proxy for GDP growth, and the other is an optimal weight to combine this proxy with national account data. After applying the method to countries with very low-quality national account data, Henderson et al. [39] demonstrated the key role of NTLs data in analyzing growth at the subnational and supranational levels. With the NTLs data, income data and population data of 748 regions across 54 countries in Africa, Mveyange [40] estimated the regional income inequality by calculate two standard measures of inequality, the Gini index and the mean log deviation (MLD) measure [41]. After presenting the empirical model, they showed that the estimated inequality index has significant and positive correlations with income-based regional inequality indicators, suggesting that NTLs are good proxies to estimate regional inequality. These results are especially meaningful in the lack of reliable and consistent subnational income data.

Cauwels et al. [42] explored the dynamics and spatial distribution of global NTLs for 160 different countries. They found that the center of light moves eastwards about 60 km per year, and there is a tendency of global centralization of light. After introducing spatial light Gini coefficients, they found a universal pattern of human settlements across different countries. Ghosh et al. [28] summarized literature that leveraged NTLs to develop a variety of alternative measures of human well-being. They introduced the application of NTLs to estimate various human well-being indicators (e.g., GDP, poverty, informal economic activity and remittances), develop the night light development index (NLDI), map the human ecological footprint, measure the electrification rates and estimate the ICT Development Index (IDI). Recently, Bennett et al. [43] summarized the methods to correlate NTLs with socioeconomic parameters including urbanization, economic activity and population. They highlighted the value of NTLs for detecting, estimating and monitoring socioeconomic dynamics.

NTLs data are successful in revealing economic activity, however, it is not effective for less developed areas due to

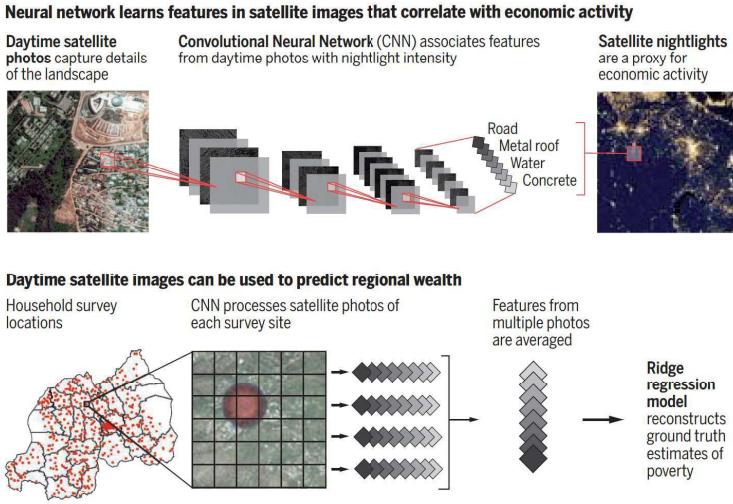


Figure 3: Predicting poverty from satellite images using Convolutional Neural Network (CNN). Figure from [21].

the uniformly dark of satellite imagery in these areas. To this point, Jean et al. [44] applied deep learning algorithms to learn the relationship between NTLs and daytime satellite imagery. The former can predict the wealth distribution while the latter contains rich information about landscape features. They employed a multi-step transfer learning approach [45] to train a convolutional neural network (CNN) [46]. In particular, a linear chain transfer learning graph was constructed. First, they transferred knowledge from the object recognition on the ImageNet (Problem 1) [47], an object classification image dataset of over 14 million images from 1000 different categories, to the prediction of NTL intensity from daytime satellite imagery (Problem 2). They chose the model trained on ImageNet as the starting CNN model [48], and then constructed the fully convolutional model. Formally, given an unrolled  $(h \times w \times d)$ -dimensional input  $x \in \mathbb{R}^{hwd}$ , the fully connected layers perform a matrix-vector product,

$$\hat{x} = f(Wx + b), \quad (1)$$

where  $W \in \mathbb{R}^{hwd}$  is a weight matrix,  $b$  is a bias term,  $f$  is a nonlinear function, and  $\hat{x} \in \mathbb{R}^k$  is the output. Then, they transferred knowledge from Problem 2 to the prediction of poverty from daytime satellite imagery (Problem 3), for which the amount of training data is limited. The illustration of the method is summarized by Blumenstock [21] (see Figure 3), and technical details are presented in the early work by Xie et al. [46]. The image features extracted from the daytime imagery can explain up to 75% of the variation in the average household asset across five African countries. Moreover, the method is able to reconstruct survey-based indicators of regional poverty with high accuracy. Using only publicly available data, the method has broad potential applications in tracking and targeting poverty in developing countries.

Other RS data and machine learning approaches can also be used in quantifying poverty-environment relationships. By applying the principal component analysis (PCA) and spatial models in the field of geostatistics, Sedda et al. [49] demonstrated the correlations between the normalized difference vegetation index (NDVI, a measure of vegetation greenness in RS [50]), intensity of poverty, and health for a large area of West Africa. They found that high NDVI is associated with low poverty and child mortality. Their results highlight the utility of satellite-based metrics for poverty analysis. With high-resolution daytime satellite imagery, the UN Global Pulse Lab Kampala built a proxy indicator for poverty based on the household's roof counting. The research project entitled "Measuring Poverty with Machine Roof Counting" [51] developed image processing software to count the roofs and identify the type of roof that a house has. Watmough et al. [52] applied a random forests approach to study the relationships between welfare and geographic metrics for over 14,000 villages in India. They found that geographic metrics account for 61% and 57% of the variation in the lowest and highest welfare quintile, respectively. These methods help estimate socioeconomic status in less developed countries where household surveys remain lacking.

### 2.1.2. Mobile phones reveal socioeconomic status

Mobile phones (MPs), serving as ubiquitous sensors, are increasingly common in developing economies. Compared to coarse-grained remote sensing, MPs are able to capture an enormous information and provide cost-effective data at the individual level, such as the frequency and timing of communication events [18, 53, 54], the traveling patterns [55], the histories of consumption and expenditure [56], and so on. With MP logs that are related to housing, education, health, etc., socioeconomic status can be inferred by employing regression models and machine learning approaches at the aggregated subnational and national levels.

To explore the relationship between MP usages and wealth in developing countries, Blumenstock et al. [56] presented a novel method that contains three steps: (1) modeling the relationship between assets and expenditures using Demographic and Health Survey (DHS) data; (2) conducting a phone-survey with a small subset of MP users to collect information on asset ownership; (3) obtaining call detail records (CDRs) for the individuals in the phone survey and creating a single dataset that use call histories to predict annual expenditures. By analyzing the data from Rwanda, they found that household expenditures are positively correlated with MP usages, mainly with the numbers of international calls, the number of different districts contacted, and the average airtime credit purchase. Airtime credit is money in MP number account, ready to spend on texts, calls and data. These results suggest that the annual expenditures of MP users can be predicted only using their anonymous phone usage data. Blumenstock and Eagle [57] later found that MP usages in Rwanda are not uniform. They provided a quantitative description about the demographic and socioeconomic structure of MP usages, for example, phone owners are considerably richer and predominantly male. Moreover, Blumenstock et al. [58] showed that Rwandans use MP network to transfer their airtime credit to those affected by disasters. In particular, transfers tend to be sent to rich individuals and between pairs of individuals with a strong history of reciprocal.

Individual MP data can be aggregated to estimate socioeconomic status at the national level. By analyzing CDRs and airtime credit purchase histories, Gutierrez et al. [59] mapped the relative income of individuals, the diversity and inequality of income, and the socioeconomic segregation for fine-grained regions in Côte d'Ivoire. In particular, they quantified the variation in purchase amounts of each user by using the Coefficient of Variation (CV),

$$CV = \sigma/\mu, \quad (2)$$

where  $\sigma$  and  $\mu$  are the standard deviation and the mean of the purchase amounts. They found that urban areas clearly stand out in diversity, showing the opportunity to obtain real-time and low-cost socioeconomic statistics. Also for Côte d'Ivoire, Smith et al. [60] demonstrated how aggregated CDRs can be mined to derive proxies of socioeconomic indicators. They found strongly negative correlations between the communication activity within a region and the multidimensional poverty index (MPI) [61], a survey-based indicator that measures a region's actual poverty. Further, they derived a linear model to estimate the poverty level using the diversity of communication. Their work suggests CDRs as an invaluable source for poverty estimation, even without the knowledge of individual behavior.

MP data from Côte d'Ivoire has also been used to explore the relations between national communication network and socioeconomic dynamics. Mao et al. [62] introduced the CallRank indicator—the PageRank centrality [63] calculated over the MP communication network—to quantify the relative importance of an area and tested the correlation between network features and socioeconomic indicators. They found that the outgoing call ratio consistently correlates with local socioeconomic statistics such as low poverty rate and high annual income. Moreover, the Gini index exhibits significant correlations with CallRank and other CDRs-based indicators. Further, to quantify the strength of the *rich-club effect* [64, 65], they measured the weighted rich-club coefficient of the MP communication network,

$$\rho^w(r) = \frac{\phi^w(r)}{\phi_{\text{null}}^w(r)}, \quad (3)$$

where  $\phi^w(r) = W_{>r} / \sum_{l=1}^{E_{>r}} w_l^{\text{rank}}$ , and  $\phi_{\text{null}}^w(r)$  corresponds to the null model generated by randomizing the original MP network while preserving its degree distribution. Here, each node has a richness parameter  $r$  as the average annual income of the region,  $E$  is the total number of links,  $E_{>r}$  is the number of links to the region,  $W_{>r}$  is the sum of the weights attached to these links, and  $w_l^{\text{rank}} \geq w_{l+1}^{\text{rank}}$  with  $l = \{1, 2, \dots, E - 1\}$  are the ranked weights of links on the network. If  $\rho^w(r) > 1$ , network shows the *rich-club effect* in comparison with the null model. The extent to which  $\rho^w(r)$  is larger than 1 indicates the strength of the *rich-club effect*. After analyzing the CDRs, Mao et al. [62] found that rich areas form rich club in MP communication, where rich areas communicate more frequently with each other.

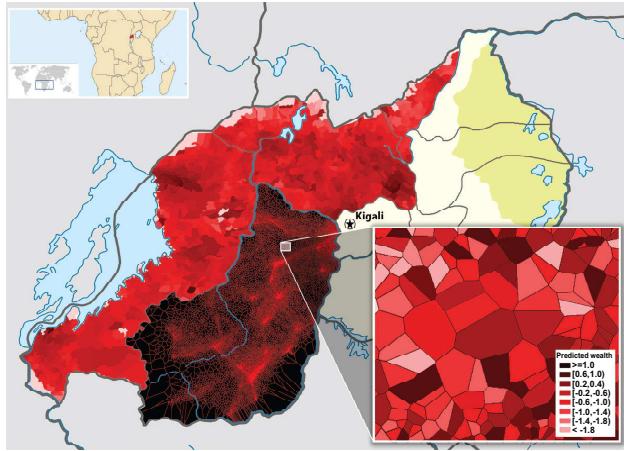


Figure 4: High-resolution map of poverty and wealth predicted from mobile phone call records of 1.5 million users in Rwanda. Figure from [66].

By analyzing anonymized records of interactions on Rwanda’s MP network and the follow-up phone surveys of some individual subscribers, Blumenstock et al. [66] predicted the wealth of MP users. They demonstrated that the predicted attributes of individuals can accurately reconstruct the distribution of the entire nation’s wealth. Specifically, they used a two-step approach in feature engineering and model selection, where the first step generates a thousand metrics from the MP data, and the second step eliminates irrelevant metrics and selects a parsimonious model using the elastic net regularization [67]. After applying this machine learning approach to analyze the survey data, they found that individual wealth can be well predicted and individuals in relative poverty can be accurately identified. Then, they generated out-of-sample predictions for 1.5 million MP users and produced the wealth map of Rwanda at a very high resolution (see Figure 4). Further, they found a strong correlation between the government “ground truth” data and the predicted wealth data after aggregating them to the district level. Their method is promising to map the distribution of wealth and other socioeconomic indicators for the full national population. Other works that leveraged MP data to infer socioeconomic status at the regional or urban levels will be introduced in the following sections.

#### 2.1.3. Combined data for better inference

Novel sources of data with a high spatial resolution have been used to provide an up-to-date indication of living conditions. For example, remote sensing (RS) data capture information about physical properties of the land, which are cost-effective but relatively coarse in urban areas. By contrast, call detail records (CDRs) from mobile phones (MPs) have high spatial resolution in urban areas but the resolution is usually insufficient in rural areas due to the sparsity of towers. Therefore, some recent works estimate socioeconomic status by combining data from different domains such as LandScan population [37], RS and MPs.

While RS-only and CDRs-only models perform comparably in mapping poverty, Steele et al. [68] demonstrated that their combination can produce better predictive maps of socioeconomic status in Bangladesh. Specifically, they employed hierarchical Bayesian geostatistical models (BGMs) [69] that combine RS data, CDRs and traditional survey-based data to map three commonly used indicators of living standards, namely, Wealth Index (WI), Progress out of Poverty Index (PPI) and reported household income (Income). The BGMs are built on the scale of the Voronoi polygons, which approximate the mobile tower coverage areas using Voronoi tessellation [70]. They applied BGMs to predict the poverty metrics (WI, PPI and Income) for each Voronoi polygon as a posterior distribution with completely modeled uncertainty around estimates. Then, they generated prediction maps with associated uncertainty using the posterior mean and standard deviation (see Figure 5). Their method using combined CDRsCRS data exhibits a better predictive power (highest  $R^2 = 0.78$ ) for the observed data than RS-only method ( $R^2 = 0.71$ ) and CDRs-only method ( $R^2 = 0.70$ ). Similarly, Njuguna and McSharry [71] built a linear model to predict MPI based on the combination of CDRs, RS and LandScan datasets in Rwanda. They extracted four meaningful features that proxy socioeconomic status from the combine dataset, specifically, nighttime lights (NTLs) per capita from RS data, mobile ownership per capita from CDRs, average daily call volume per phone from CDRs, and population density from LandScan data.

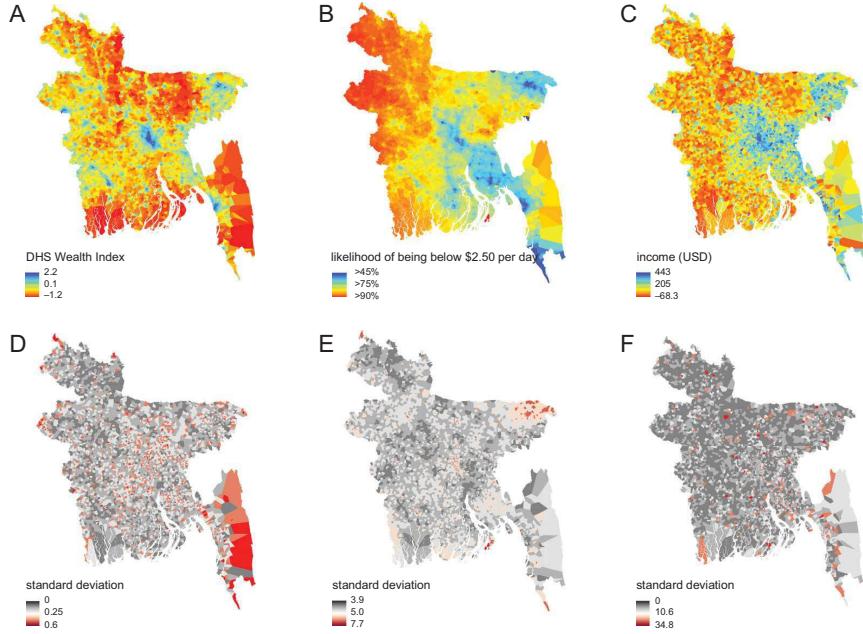


Figure 5: Maps of predicted living standards based on call detail records (CDRs) and remote sensing (RS) data in Bangladesh. Mean wealth index (A) with uncertainty (D); mean likelihood being below \$2.50/day (B) with uncertainty (E); and mean income (C) with uncertainty (F). The maps show the posterior mean and standard deviation from CDR-RS models for the WI and income data (A,C), and the RS model for the PPI (B). Red color indicates poorer areas in prediction maps, and higher error in uncertainty maps. Figure from [68].

They proposed a simple linear regression model using the four features to predict MPI, as

$$\log(\text{MPI}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4, \quad (4)$$

where  $x_i$  stands for the value of the corresponding feature. This model can explain 76% of the variance in MPI across 295 sectors in Rwanda. These results suggest that combination of multiple data sources can yield socioeconomic estimates at a high spatial resolution.

Extract from digital and physical commodities can provide rich information about socioeconomic status, and thus proxy indicators can be built by leveraging these novel data sources. For example, United Nations Global Pulse launched the project entitled “Building Proxy Indicators of National Wellbeing with Postal Data” [72], which investigates the potential of using the international postal flow network to approximate indicators of countries’ socioeconomic profiles. The project collected 14 million electronic postal records of 187 countries from 2010 to 2014. The dataset covers 680,000 post offices and forms the world’s largest postal network. Results show that indicators gathered from the postal network correlate well with fourteen widely used socioeconomic indicators such as GDP and Human Development Index (HDI). This work demonstrates that structural features of world flow networks can be used to produce proxy indicators of socioeconomic status.

Meanwhile, Hristova et al. [73] examined how digital traces and the network structure can reveal the socioeconomic profiles of different countries. They measured the position of each country in six different global networks (trade, postal, migration, international flights, IP and digital communications) and built proxies for a number of socioeconomic indicators including GDP per capita and HDI ranking and other twelve indicators. In particular, they applied the multilayer network model [74] to characterize the strength of these international ties, where six networks representing six types of international ties are considered as six layers of the multiplex network with each pair of nodes possibly having one relationship in each layer. Formally, the multiplex network [75] is denoted as

$$\mathcal{M} = \left\{ G^1(V^1, E^1), \dots, G^a(V^a, E^a), \dots, G^m(V^m, E^m) \right\}, \quad (5)$$

where each layer contains a set of edges  $E$  and a set of nodes  $V$ , and  $m = 6$  is the total number of networks. The

multiplex neighborhood of a node  $i$  is defined as the union of its neighborhoods on each layer:

$$N_{\mathcal{M}}(i) = \left\{ N_{\alpha}(i) \cup N_{\beta}(i) \dots \cup N_m(i) \right\}, \quad (6)$$

where  $N_{\alpha}(i)$  is the neighbourhood of node  $i$  in layer  $\alpha$ . The global multiplex degree of node  $i$  is defined as  $k^{\text{glob}}(i) = |N_{\mathcal{M}}(i)|$ , and the weighted global multiplex degree is defined as

$$k_w^{\text{glob}}(i) = \sum_{j \in N_{\mathcal{M}}(i)} \sum_{G \in \mathcal{M}} \frac{e_{ji}}{n \times m}, \quad (7)$$

where  $n$  is the total number of nodes. The network metrics have predictability to several socioeconomic indicators. The global multiplex degree is the best-performing degree in terms of consistently high performance across all fourteen indicators. In particular, the global degree exhibits the most highly negative correlation with the HDI ranking (Spearman's rank correlation  $\rho \approx 0.8$ ). These results show that a nation's socioeconomic proxy indicators can be constructed based on different global networks after combining the data from multiple sources.

## 2.2. Economic complexity and fitness of nations

Understanding how economies develop to prosperity is a long-standing challenge in economics. In traditional literature, as an aggregated monetary indicator, GDP has been widely used to identify the stages of economic development of countries. Recently, a novel index named economic complexity has been proposed as the root in the gaps of economic development. In particular, the new stream of literature introduce a variety of non-monetary metrics based on international trade networks to quantitatively assess a country's potential for future economic growth. In this section, we will briefly introduce recent works on economic complexity index, fitness index, and some variant indices, as well as their applications to predict world economic development.

### 2.2.1. Product space and economic complexity

Economic development has been traditionally measured by aggregated variables like GDP, however, such averages can not capture the increasing diversity that is associated with economic development. An insight raised recently is that the mix and diversity of products and industries are highly suggestive to economic growth. Hausmann et al. [76] introduced the level of sophistication—the income level of a country's exports—to the characterization of products and demonstrated that it can predict subsequent economic growth. Specifically, they first construct an index called PRODY, which represents the income level associated with a product. The PRODY index for product  $p$  is given by

$$\text{PRODY}_p = \sum_c \frac{(x_{cp}/X_c)}{\sum_{c'}(x_{c'p}/X_{c'})} Y_c, \quad (8)$$

where  $x_{cp}$  is the total export of product  $p$  by country  $c$ ,  $X_c = \sum_p x_{cp}$  is the total export of country  $c$ , and  $Y_c$  is the GDP per capita of country  $c$ . Indeed, the PRODY index is a weighted average of the per capita GDPs of countries exporting a given product. Then, they construct the PRODY index, which represents the income level associated with a country's export basket. The PRODY index for country  $c$  is given by

$$\text{EXPY}_c = \sum_p \left( \frac{x_{cp}}{X_c} \right) \text{PRODY}_p. \quad (9)$$

Indeed, the PRODY index is a weighted average of the PRODY for the country, where the weights are the shares of the products in the total exports of the country. After analyzing the international trade data covering over 5,000 products and 124 countries, Hausmann et al. [76] found that countries with high initial sophistication of export baskets (EXPY) tend to perform better in subsequent economic growth. These results suggest that countries have economically meaningful differences in the specialization patterns of exporting baskets, and countries export more sophisticated products are likely to grow more rapidly.

Later, Hidalgo et al. [77] illuminated this viewpoint through analyzing the network of relatedness between products, named product space, which is built based on the international trade data. Products are considered to have high

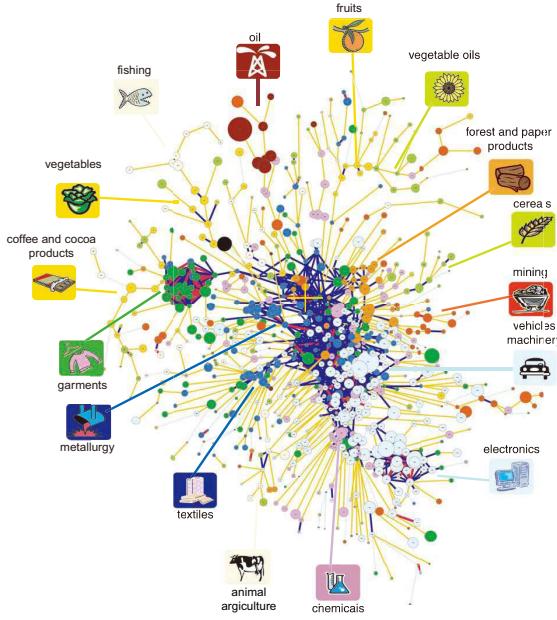


Figure 6: The network representation of the product space built on the international trade data. Links are color coded with their proximity value. The sizes of the nodes are proportional to world trade, and their colors are chosen according to the product classification. Figure after [77].

relatedness if they have a high probability to be co-exported by many countries in the international trade. Formally, the proximity between products  $i$  and  $j$  is defined as

$$\phi_{ij} = \min \{P(\text{RCA}_x_i | \text{RCA}_x_j), P(\text{RCA}_x_j | \text{RCA}_x_i)\}, \quad (10)$$

where  $P(\text{RCA}_x_i | \text{RCA}_x_j)$  is the conditional probability that country  $x$  is a significant exporter of product  $i$  given that it has been a significant exporter of product  $j$ . The significant exporter of a product is identified by the revealed comparative advantage (RCA) [78]. The RCA value is defined as the share of product  $p$  in the export basket of country  $c$  to the share of product  $p$  in the world trade. Specifically, the  $\text{RCA}_{cp}$  of country  $c$  in product  $p$  is defined by

$$\text{RCA}_{cp} = \frac{x_{cp}}{\sum_{p'} x_{cp'}} \left/ \frac{\sum_{c'} x_{c'p}}{\sum_{p'} \sum_{c'} x_{c'p'}} \right., \quad (11)$$

where  $x_{cp}$  is the total export of product  $p$  by country  $c$ . If  $\text{RCA}_{cp} \geq 1$ , country  $c$  is a significant exporter of product  $p$ . Larger proximity  $\phi_{ij}$  means higher relatedness between products  $i$  and  $j$ . Based on the proximity measure, the product space is generated and visualized (see Figure 6). It can be seen that the product space has a core-periphery structure with more-sophisticated products locating in the core and less-sophisticated products occupying the periphery (see Ref. [79, 80] for the definition of core-periphery structure in networks). Richer and poorer countries tend to export products that are located in the core and periphery, respectively. More significantly, countries move through the “product space” by developing products that are related to what they currently have. These results provide explanations to the fact that economic development a path-dependent process [81] and not all countries face the same opportunities in development.

In particular, it is hard for poor countries to move toward new products with high sophistication since these countries tend to occupy the peripheries of the product space with current exports of less-sophisticated products. Using the concept of product space to explore the international trade data, Abdon and Felipe [82] studied the opportunity for economic growth and structural transformation of Sub-Saharan Africa (SSA) countries. They found that the majority of SSA countries are trapped in the export of products that are unsophisticated, standard and poorly connected in the product space. This makes the structural transformation of a region being particularly difficult, because the nearby products are in the periphery and the current capabilities are not enough to jump into more sophisticated products. To

solve this problem, governments must implement policies and provide public inputs that can give incentives for the private sector to invest in the more sophisticated activities.

Further, Hidalgo and Hausmann [83] quantified the economic complexity of nations based on international trade data and demonstrated its central role in a country's economic development. In particular, they proposed the Method of Reflections (MR) to characterize the structure of "country-product" bipartite network and showed that the variables produced by the MR method can be interpreted as indicators of economic complexity. Formally, the bipartite network can be represented by an adjacency matrix  $M_{cp}$ , where  $M_{cp} = 1$  if country  $c$  is a significant exporter ( $RCA_{cp} \geq 1$ ) of product  $p$ , and  $M_{cp} = 0$  if otherwise. The economic complexity index (ECI) of country  $c$  is then defined as

$$\text{ECI}_c = \frac{K_c - \langle \vec{K} \rangle}{\text{std}(\vec{K})} = \frac{N^2 K_c - N \sum_c K_c}{\sqrt{N \sum_c (NK_c - \sum_c K_c)^2}}, \quad (12)$$

where  $N$  is the number of countries,  $\langle \cdot \rangle$  and  $\text{std}(\cdot)$  are functions of mean and stand deviation that operate on the elements of vector  $\vec{K}$ , and  $\vec{K}$  is the eigenvector associated with the 2nd largest eigenvalue of the matrix

$$\tilde{M}_{cc'} = \sum_p \frac{M_{cp} M_{c'p}}{k_{c,0} k_{p,0}}. \quad (13)$$

Indeed, the matrix  $\tilde{M}_{cc'}$  is defined through a set of linear iterative equations by connecting countries who have similar products, weighted by the inverse of the ubiquity of product ( $k_{p,0} = \sum_c M_{cp}$ ) and normalized by the diversity of country ( $k_{c,0} = \sum_p M_{cp}$ ). Formally, putting the equation  $k_{p,N} = \sum_c M_{cp} k_{c,N-1}/k_{p,0}$  (the average ubiquity of product) into the equation  $k_{c,N} = \sum_p M_{cp} k_{p,N-1}/k_{c,0}$  (the average diversity of country) can generate the equation

$$k_{c,N} = \sum_{c'} k_{c',N-2} \sum_p \frac{M_{cp} M_{c'p}}{k_{c,0} k_{p,0}}, \quad (14)$$

where  $N \geq 2$  is the number of iteration. The economic complexity of country  $c$  is given by  $\text{ECI}_c = \sum_{c'} \tilde{M}_{cc'} \text{ECI}_{c'}$ , where  $\text{ECI}_{c'}$  is country  $c'$ 's complexity in the previous iteration step. For more mathematical details, readers are encouraged to read the book on economic complexity wrote by Hausmann et al. [84]. Empirical results showed that countries' ECIs are highly correlated with their income levels are predictive of their future growth. Indeed, economic development is a process that requires acquiring more complex sets of capabilities to move towards new activities associated with higher levels of productivity. Therefore, efforts should focus on generating the conditions that allow complexity to emerge, so that sustained growth and prosperity in economic development will appear.

From a network perspective, uncovering the characteristics of the "country-product" bipartite network is very important for understanding economic development. Hausmann and Hidalgo [85] proposed an analytic framework to account for the nature of the bipartite network structure. They found that countries differ in their product diversification and in the ubiquity of their exported products. Countries with more capabilities are able to produce less ubiquitous products. This logic explains the negative relationship between the diversification of countries and the average ubiquity of the products that they produce. Later, Bustos et al. [86] studied the presence and absence of industries in international and domestic economies. They found that "country-product" bipartite networks are significantly nested [87], and the dynamics of nestedness can predict the evolution of industrial ecosystems (see Refs. [88, 89] for details on nestedness in networks). Moreover, the nestedness tends to be constant over time, making the pattern of industrial appearances predictable. Felipe et al. [90] applied MR to rank 5107 products and 124 countries in the international trade. They found that countries' export shares of products of different complexity vary with the level of their income per capita. Specifically, export shares of the most complex products increase with income, while the export share of the less complex products decrease with income. Moreover, MR can distinguish products that require more complex or simpler capabilities, and the complexity rankings of countries exhibit a high correlation with their technological capabilities.

### 2.2.2. Fitness index and economic dynamics

The Fitness index employs a statistical approach to define a new set of metrics to quantify the fitness of countries and the complexity of products through coupled nonlinear maps. Based on the analysis of the "country-product"

bipartite networks of international trade, Caldarelli et al. [91] proposed a new method based on biased Markov chain process to rank countries in a more conceptually consistent way, where a two-parameter bias is used to account for the bipartite network structure. Formally, the Markov process is given by

$$\begin{cases} w_c^{(N+1)}(\alpha, \beta) = \sum_p G_{cp}(\beta) w_p^{(N)}(\alpha, \beta) \\ w_p^{(N+1)}(\alpha, \beta) = \sum_c G_{pc}(\alpha) w_c^{(N)}(\alpha, \beta) \end{cases}, \quad (15)$$

where  $w_c$  is the fitness of country  $c$ ,  $w_p$  is the complexity of product  $p$ ,  $N$  is the interaction step, and  $G$  is the Markov transition matrix given by

$$\begin{cases} G_{cp}(\beta) = \frac{M_{cp} k_c^{-\beta}}{\sum_{c'} M_{c'p} k_{c'}^{-\beta}} \\ G_{pc}(\alpha) = \frac{M_{cp} k_p^{-\alpha}}{\sum_{p'} M_{cp'} k_{p'}^{-\alpha}} \end{cases}, \quad (16)$$

where  $\alpha$  and  $\beta$  are free parameters. In a vectorial formalism, country  $c$ 's fitness is  $\mathbf{w}_c^{(N+1)}(\alpha, \beta) = T(\alpha, \beta) \mathbf{w}_c^{(N)}(\alpha, \beta)$ , where the ergodic stochastic matrix  $T$  is defined as  $T_{cc'}(\alpha, \beta) = \sum_p G_{cp}(\beta) G_{pc'}(\alpha)$ . The complexity of product  $p$  is  $\mathbf{w}_p^{(N+1)}(\alpha, \beta) = S(\alpha, \beta) \mathbf{w}_p^{(N)}(\alpha, \beta)$ , where the ergodic stochastic matrix  $S$  is  $S_{pp'}(\alpha, \beta) = \sum_c G_{pc}(\alpha) G_{cp'}(\beta)$  (see Ref. [91] for mathematical details). After analyzing these equations, Caldarelli et al. [91] revealed a strongly nonlinear entanglement between the diversification of a country and the ubiquity of its products in determining the competitiveness of countries and the complexity of products. In particular, having more-sophisticated products in the portfolio contributes more to the competitiveness of a country than having many less-sophisticated products.

Moving forward, Tacchella et al. [92] developed a so-called Fitness-Complexity Method (FCM) using coupled nonlinear maps, whose fixed point can define new metrics for the fitness of countries and the complexity of products. In their iterative algorithm, fitness of countries and complexity of products interact in a nonlinear and self-consistent mathematical way. Specifically, the fitness of a country is proportional to the number of its products weighted by their complexity. In turn, the complexity of a product is inversely proportional to the number of countries exporting it weighted by the inverse of their fitness (similar methods have also been proposed for search engine [93] and online reputation systems [94]). Formally, the coupling between the fitness  $F_c$  of country  $c$  and the complexity  $Q_p$  of product  $p$  is given by the nonlinear iterative scheme:

$$\begin{cases} \tilde{F}_c^{(N)} = \sum_i M_{cp} Q_p^{(N-1)} \\ \tilde{Q}_p^{(N)} = \frac{1}{\sum_c M_{cp} \frac{1}{F_c^{(N-1)}}} \end{cases}, \quad (17)$$

where  $\tilde{F}_c^{(N)}$  and  $\tilde{Q}_p^{(N)}$  are respectively normalized in each step by  $F_c^{(N)} = \tilde{F}_c^{(N)} / \langle \tilde{F}_c^{(N)} \rangle$  and  $Q_p^{(N)} = \tilde{Q}_p^{(N)} / \langle \tilde{Q}_p^{(N)} \rangle$ , given the initial condition  $F_c^{(0)} = 1$  and  $Q_p^{(0)} = 1$ . The nonlinear iteration goes until the stationary state is reached (see Ref. [95] for the convergence property), in which  $F$  reflects the fitness of countries and  $Q$  reflects the complexity of products. Indeed, FCM is based on the idea that (i) a diversified country gives limited information on the complexity of products, and (ii) a poorly diversified country tends to have a specific product of a low level sophistication. Therefore, a nonlinear iteration is needed to bound the complexity of industries by the fitness of the less competitive provinces having them. After applied to the international trade data, FCM performs better than MR in capturing the bipartite network structure, in defining an effective non-monetary metric for economic complexity, and in quantifying a country's potential for growth.

Meanwhile, Cristelli et al. [96] argued that nonlinear dependence is the fundamental element and the nonlinear approach is consistent with the structure of the unweighted “country-product” bipartite network. Moreover, they analyzed the case of including weights in the matrix  $M_{cp}$  through  $M_{cp} = x_{cp} / \sum_{c'} x_{c'p}$ , where  $x_{cp}$  is the total export of product  $p$  by country  $c$ . After comparing MR and FCM in both economic and mathematical aspects, they found that FCM is more conceptually consistent and well-grounded from an economic point of view. Taking into account the triangular structure of the bipartite network, Tacchella et al. [97] discussed how to define suitable non-monetary metrics for both the complexity of products and the diversification of countries. In particular, they argued the conceptual flaws of MR by using three toy models and demonstrated that FCM is able to grasp the level of competitiveness of a country by defining the simplest metrics that seem to be consistent with the triangular-like pattern.

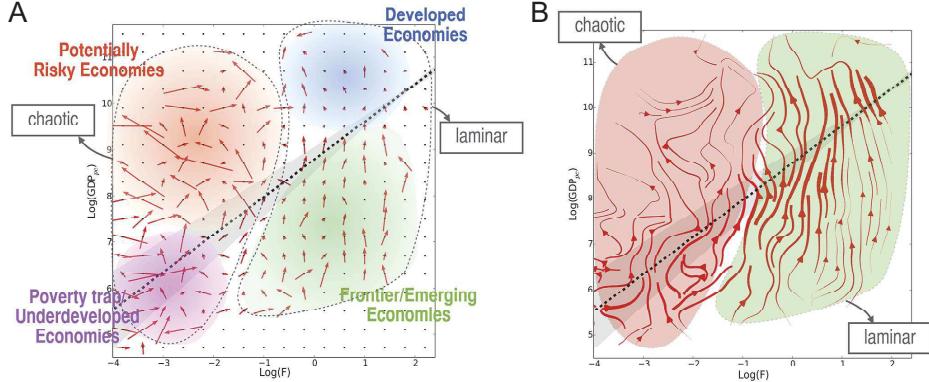


Figure 7: The heterogeneous dynamics of countries in the Fitness-GDPpc plane. (a) A finer coarse graining of the dynamics highlights two regimes. One regime is the laminar region (right), where fitness is the driving force of the growth, and the evolution of countries in this region is highly predictable. The other regime is the chaotic region (left), where the issues are very close to the problems of predictability for dynamical systems and to develop a predictive scheme using tools like regressions is no more appropriate. (b) The continuous interpolation of the coarse grained dynamics. The predictability in the two regimes, laminar and chaotic, is better illustrated. Figure from [98].

This branch of studies has provided new perspectives to cast economic prediction into the conceptual scheme of forecasting the evolution of a dynamical system, for example, weather dynamics. Cristelli et al. [98] compared the non-monetary metrics, in particular the fitness of countries, with their monetary figures, say GDPpc. They showed that FCM is able to quantify the hidden growth potential of countries. More interestingly, they demonstrated that the pattern of countries' evolution in the Fitness-GDPpc plane is strongly heterogeneous with two regimes of very different predictability features (see Figure 7). Specifically, there is a strongly predictable area of economic development, named the laminar regime, while the predictability is low in the so-called chaotic regime. Two kinds of evolution patterns can be observed in the laminar regime, where emerging economies develop rapidly and developed economies enjoy stable growth. In the chaotic regime, the dynamics of countries are highly diverse and unstable, leading to the difficulty in predicting the economic development. In this case, tools like regressions are no more appropriate in developing a predictive scheme.

To address this issue, Cristelli et al. [98] defined a selective predictability scheme to assess future evolution of countries by resembling the method of analogues [99], which was developed to predict the evolution of a dynamical system given the knowledge of the past but without the laws of motion. The framework provides insights to the regime-dependent economic predictability and opens new paths to economic forecasting. Recently, Tacchella et al. [100] applied this scheme to predict the five-year GDP growth. In the Fitness-GDPpc plane, they repeatedly sampled analogues with a Gaussian kernel (centred on the present state of a country) and performed a bootstrap of previously observed evolution (weighted by the distance of the analogues starting points), resulting in the global distribution of possible outcomes. They further refined the forecast by taking into account the strong self-correlation of GDP growth. Specifically, the forecast based on the global distribution is combined with the forecast that assumes a past five-year growth by a certain weighted averaging. This scheme outperforms the International Monetary Fund (IMF) five-year GDPpc forecast [101] by more than 25% in accuracy. Moreover, the method's forecasting errors are predictable and not correlated with IMF errors, showing its complementarity to traditional approaches.

### 2.2.3. Variant indices and development analysis

Many recent studies have highlighted the importance of complexity and capabilities in economic development. The pioneering work by Hidalgo and Hausmann [83] introduced MR to extract the competitiveness of countries and the complexity of product from the “country-product” bipartite networks with the assumption that there are linear interactions between the two metrics. Tacchella et al. [92, 97] proposed FCM and emphasized the necessary of nonlinear coupling between the fitness of countries and the complexity of products. Mariani et al. [102] quantitatively compared the ability of MR and FCM in ranking countries and products by their importance in networks. Based on the international trade data of 132 countries and 723 products, they found that FCM outperforms MR in ranking both products and countries. In particular, FCM captures the nestedness of the bipartite network and ranks nodes better by

their importance.

Mariani et al. [102] proposed a modified FCM (MFCM for short), in which the nonlinear coupling is governed by a tunable parameter. By adjusting the parameter, we can find a better tradeoff between the favor on countries with diversified exports and the penalization on products with a large number of exporting countries. Formally, MFCM is defined by the equations

$$\begin{cases} \tilde{F}_c^{(N)}(\gamma) = \sum_i M_{cp} Q_p^{(N-1)} \\ \tilde{Q}_p^{(N)}(\gamma) = \left[ \sum_c M_{cp} (F_c^{(N-1)})^{-\gamma} \right]^{-1/\gamma}, \end{cases} \quad (18)$$

where  $\gamma$  is the tunable parameter. When  $\gamma = 1$ , MFCM degenerates to FCM. The correlation between the product complexity  $\tilde{Q}(\gamma)$  and the product ubiquity  $k_{p,0}$  decreases with the increase of  $\gamma$ . When  $\gamma \gtrsim 2$ , the ranking of product complexity  $\tilde{Q}(\gamma)$  by MFCM is perfectly correlated with that by FCM, however the ranking is volatile (very sensitive to noise). For this reason, MFCM with larger  $\gamma$  can only be applied to high-quality data instead of noisy data. When input data is reliable, MFCM is able to produce better rankings of countries and products.

Wu et al. [103] showed some rigorous mathematical properties of the fitness-complexity metric for nested networks. They introduced a simpler variant of FCM, named Minimal Extremal Metric (MEM), where the complexity of a product  $p$  is equal to the fitness of the least-fit country that exports it. Formally, MEM defines the fitness of country  $c$  and the complexity of product  $p$  by

$$\begin{cases} \tilde{F}_c^{(N)} = \sum_i M_{cp} Q_p^{(N-1)} \\ F_c^{(N)} = \tilde{F}_c^{(N)} / \langle \tilde{F}_c^{(N)} \rangle \\ Q_p^{(N)} = \min_{i: M_{cp}=1} F_c^{(N)} \end{cases} \quad (19)$$

Obviously, in MEM, only the fitness of the least-fit country  $\min_{i: M_{cp}=1} F_c^{(N)}$  contributes to the product complexity  $Q_p^{(N)}$ . In the limit  $\gamma \rightarrow \infty$ , MEM is a special case of MFCM. Results based on the analysis of the international trade data show that MEM can reproduce the nested structure of the “country-product” bipartite network but it is highly sensitive to noise in data.

Morrison et al. [104] provided both theoretical and numerical evidence for the intrinsic instability in the nonlinear map employed by FCM. Using the preferential attachment model (see Refs. [105, 106]) and two real-world datasets (trade and patent), they showed that FCM is unstable to even small perturbations in the network, while MR does not suffer from this problem. That is because the nonlinear iterative approach in FCM amplifies the effects of countries with low fitness on the complexity of a product and highlights economies producing exclusive niche products, which are produced by a very few countries but not necessarily the most sophisticated. Adding a product exported by only a single country may lead to a global reorganization of the fitness landscape. Therefore, FCM has a serious problem when applied to dynamic economical systems with new products entering markets.

With new methodologies, attentions have been paid to better understand economic development, innovation and industrialization. Based on the international trade data, Zaccaria et al. [107] built a hierarchically directed network by measuring the taxonomy of products through computing the excess frequency of co-occurrence of two products comparing to the random binomial case. Formally, the taxonomy between products  $p$  and  $p'$  is defined by projecting the “country-product” matrix  $M_{cp}$  to a unipartite space as (similar to [108])

$$B_{pp'} = \frac{1}{\max \{ \sum_c M_{cp}, \sum_c M_{cp'} \}} \sum_c \frac{M_{cp} M_{cp'}}{\sum_p M_{cp}}. \quad (20)$$

The taxonomy network presents the temporal connections between products and suggests the most relevant products for the development of countries. Indeed, the structure of the taxonomy network is suggestive to the potential growth of countries. Later, Saracco et al. [109] proposed a dynamical network approach to model the process of country’s innovation and competition on the evolution of the export baskets. Their dynamical model can accurately reproduce the main features observed in the evolution of the “country-product” bipartite network. Moreover, their model suggests that countries can follow different paths in the “product space” [77, 107] to gradually diversify their export baskets.

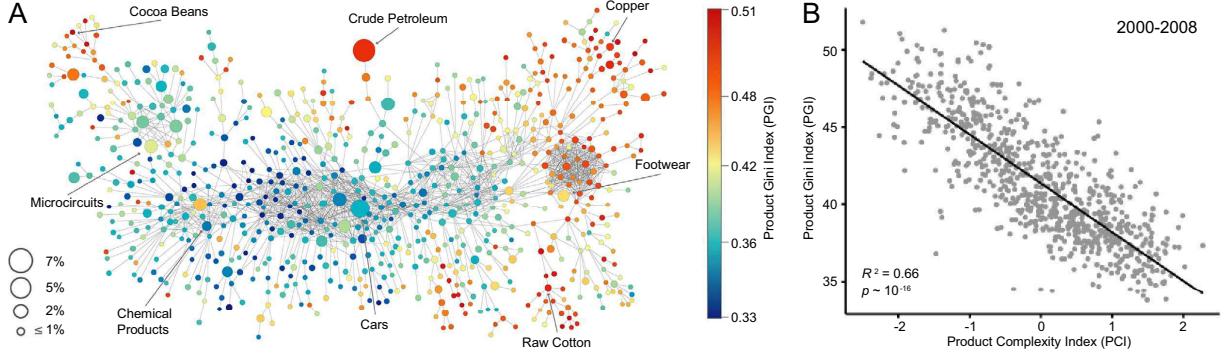


Figure 8: The product space and income inequality. (A) In the product space, nodes are colored according to the Product Gini Index (PGI) as measured during 1995-2008. The sizes of nodes are proportional to the volume of the international trade during 2000-2008. The networks are based on a proximity matrix representing 775 SITC-4 product classes exported during 1963-2008. The link strength (proximity) is based on the conditional probability that the products are co-exported. (B) The relationship between the Product Complexity Index (PCI) and the Product Gini Index (PGI) in the 2000-2008. Figure from [112].

Focusing on the time evolution of trade volume, average complexity and competitiveness, Zaccaria et al. [110] compared the exports of different sectors in Netherlands. They found that high-tech related sectors have high average complexity but low competitiveness, while sectors heavily relying on raw materials have a low complexity but high competitiveness such as Energy and Horticulture sectors. Indeed, not only products but also services are important in explaining economic stability and predicting future growth. Stojkoski et al. [111] found that services have in general higher economic complexity than products. The sophistication and diversification of service exports can provide an additional route for economic growth in both developing and developed countries. Countries that are not able to diversify service portfolio may face diminishing growth prospects.

Hartmann et al. [112] found that countries exporting more complex products have lower levels of income inequality. In particular, economic complexity index (ECI) outperforms GDP in explaining income inequality. Based on the international trade data, they calculated the Product Complexity Index (PCI) using the method proposed by Hidalgo and Hausmann [83]. Further, they estimated the level of income inequality associated with products by introducing the Product Gini Index (PGI), which is a weighted average of the Gini coefficients of the countries that export a product (see Figure 8A for the PGIs of products in the product space). There is a strong and negative correlation between PCI and PGI, showing that sophisticated products tend to have low levels of inequality (see Figure 8B). Moreover, countries with high (low) level of ECI are more likely to specialize in high-PCI (low-PCI) products, suggesting that the productive structure of a country may condition its range of income inequality. Recently, Mealy et al. [113] interpreted economic complexity metrics by showing that ECI and PCI are equivalent to a spectral clustering algorithm, which divides a similarity network into two parts. Moreover, these measures are closely related to many dimensionality reduction methods such as correspondence analysis and diffusion maps. Their findings shed some new light on the empirical success of ECI and PCI in explaining specialization patterns of countries in economic growth.

Pugliese et al. [114] analyzed the role of complexity in economic development and found that economies with differentiated products face a lower barrier in the transition towards industrialization. They extended the concept of poverty trap to include the two factors of economic complexity and GDPpc (see also Ref. [98]). They defined an index of development and industrialization, named Complex Index of Relative Development (CIRD), by the equation:

$$\text{CIRD}_{c,t} = \beta(\log F_{c,t}) + (1 - \beta)\log(GDPpc_{c,t}), \quad (21)$$

where  $F_{c,t}$  is the fitness of country  $c$  at time  $t$ ,  $GDPpc_{c,t}$  is the GDPpc of country  $c$  at time  $t$ , and  $\beta$  is a tunable parameter. The use of the CIRD index allows to study development as a monodimensional process. In particular,  $\text{CIRD}_{c,t} \approx -2$  is a threshold for countries to exit the poverty trap, and the increase of the input growth reaches its maximum at this critical point. The CIRD index facilitates our understanding of industrialization dynamics and is helpful for development analysis. Sbardella et al. [115] analyzed the relationship between wage inequality and industrialization using fitness and GDPpc. They found that movement of wage inequality along with the industrialization

follows a longitudinally persistent pattern. This finding is comparable to theories proposed by Kuznets [22], who hypothesized that countries with an average level of development suffer the highest levels of wage inequality.

Along with the literature, some online platforms have been developed and launched to help understand the evolution of countries' productive structures and economic development. For example, Simoes and Hidalgo [116] launched a data visualization site, named Observatory of Economic Complexity (OEC) (<https://atlas.media.mit.edu>). The OEC combines a number of international trade datasets and serves more than millions of interactive visualizations including imports and exports, origins and destinations, product space, economic complexity rankings based on MR, income inequality, and so on. Meanwhile, the GROWTHCOM Project launched a data platform (<http://www.growthcom.eu>), which provides visualization tools of the product network [107] and the countries' trajectories in the fitness-GDPpc plane [98].

### 2.3. Spatial demography and culture evolution

High resolution and near real-time data from new sources like remote sensing (RS), mobile phone (MP) and social media (SM) are complementary to traditional costly data with a long-time delay in inferring population distributions and demographics. Moreover, these so-called socioeconomic big data, together with methods from interdisciplinary fields including statistical physics and computer sciences, have been used to predict international migration and quantify world culture evolution. In this subsection, we will briefly introduce some methods using new data sources to map world population, estimate international migration and study culture evolution.

#### 2.3.1. World population distribution

Knowing the spatial distribution of population on earth is critical for many socioeconomic applications such as accurate environmental impact assessments, human health adaptive strategies and disease burden estimation [117]. Developed countries have substantial resources to create accurate and contemporary population datasets with high spatial resolution [118], however, relevant data are often scarce, outdated and unreliable in low-income countries due to economic constraints. In addition, acquiring census data in a timely and accurate manner is very difficult due to the rapid change of population and some administrative challenges. As a result, our knowledge of population distribution in many areas of the world remains poor thus far. Fortunately, technologies developed during the past decades have opened new ways for us to estimate and map world population distribution in a more timely manner and with a relatively lower cost.

Some large-area gridded world population distribution datasets have been built based on multiple data resources. Tobler et al. [119] developed the first version of Gridded Population of the World (GPW) database by transforming population counts from census units to a grid. The Global Rural Urban Mapping Project (GRUMP) utilizes higher resolution inputs and renders outputs at a 30 arc-second resolution (approximately 1km). In addition to census data, spatial covariate datasets are also used to estimate populations. For example, the LandScan Global Population Project [37] produced the world-wide 1998 LandScan population database at a 30 arc-second resolution based on the land cover database derived from satellite imagery and urban area vector data [120]. Tatem et al. [121] produced the 100m gridded population map by combining land cover information and census data under the Malaria Atlas Project. The semi-automated population distribution mapping at unprecedented spatial resolution produces more accurate results at a spatial resolution of about 100m in East Africa.

Cheriyadat et al. [122] generated human settlement maps based on high-resolution satellite imagery. Their algorithm employed gray level co-occurrence matrices [123] to generate texture and edge patterns from satellite imagery that are useful in urban land cover classification. Liao et al. [124] presented a high-accuracy population mapping method that integrates genetic programming (GP) [125] and genetic algorithms (GA) [126] with geographic information systems (GIS). Specifically, they applied GIS to identify relevant factors (e.g., land-cover types and transport infrastructure) and use GP and GA to transform census data to population grids. Deng et al. [127] estimated small-area population by incorporating GIS, remote sensing (RS) and demographic data into a popular demographic model. They demonstrated that the derived spatial factors can significantly improve the accuracy of small-area population estimation.

Gaughan et al. [128] constructed an accurate and high-resolution population distribution dataset for Southeast Asia. They modeled population distributions for 2010 and 2015 by combining satellite-derived settlement maps, land cover information, and ancillary datasets on infrastructure. Stevens et al. [129] presented a new semi-automated

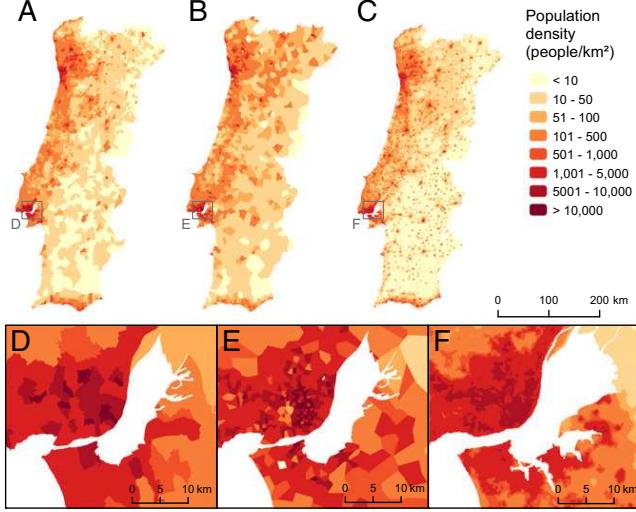


Figure 9: Comparison of predicted population density datasets with baseline data for mainland Portugal. (A) Population density derived from the national census. (B) Population estimated by the mobile phone method. (C) Population density estimated by the remotely sensing method. (D-F) Close-ups around the capital city Lisbon. Figure from [135].

dasymetric modeling approach, where RS and geospatial data are combined to model the dasymetric weights and the random forest model is used to generate a gridded prediction of population density at about 100m resolution. Patel et al. [130] presented a novel method to map multitemporal settlement and population from Landsat imagery using Google Earth Engine, which is an online environmental data monitoring platform that provides analysis capabilities on Landsat data by leveraging cloud computing services. They demonstrated that the integration of GEE-derived urban extents improves the quality of population mapping.

Spatial covariates derived from satellite imagery and land cover are typically static in nature and are not direct measures of people's presence on earth [118]. Thanks to the rapid adoption of Internet and mobile devices in developing countries, there is a great potential of using digital records to do population mapping. For example, call detail records (CDRs) can overcome many limitations of census-based data since MPs have a high penetration rate across the world. For urban areas, Pulselli et al. [131] developed a technique to monitor population density in real time based on MP chatting, given that the intensity of activity in the area covered by an antenna is proportional to the number of MP users. Based on MP location data, Dan and He [132] proposed a dynamic distribution model to estimate urban population density using an improved K-means clustering algorithm [133]. Kang et al. [134] discussed several fundamental issues on using CDRs to estimate population distributions. After analyzing the CDRs of nearly two million MP subscribers, they found that the number of calls other than the total daily call volume serves as a good estimator of population distribution.

Recently, using both RS and MP data, Deville et al. [135] produced spatially and temporally explicit estimations of population densities at national scales (see Figure 9). Based on over one billion CDRs from Portugal and France, they estimated the population density of an administrative unit  $c_i$  using a two-step method that relies on the density  $\sigma_{v_j}$  of MP users, where  $v_j$  is the Voronoi polygon [70] associated with tower  $j$ . The nighttime density  $\sigma_{c_i}$  for unit  $c_i$  is calculated by

$$\sigma_{c_i} = \frac{1}{A_{c_i}} \sum_{v_j} \sigma_{v_j} A_{(c_i \cap v_j)}, \quad (22)$$

where  $A_{c_i}$  is the area of unit  $c_i$ , and  $A_{(c_i \cap v_j)}$  is the intersection area of unit  $c_i$  and the Voronoi polygon  $v_j$ . The density  $\sigma_{c_i}$  is compared with the census-derived population densities  $\rho_{c_i}$  through

$$\rho_c = \alpha \sigma_c^\beta, \quad (23)$$

where  $\rho_c = [\rho_{c_1}, \rho_{c_2}, \dots, \rho_{c_n}]$  and  $\sigma_c = [\sigma_{c_1}, \sigma_{c_2}, \dots, \sigma_{c_n}]$ . By transforming Eq. (23) to  $\log(\rho_c) = \log(\alpha) + \beta \log(\sigma_c)$ , the two parameters  $\alpha$  and  $\beta$  can be fitted by a linear regression on training data. Further, they combined the MP

method with the RS method proposed by Stevens et al. [129], who used the random forest model to generate gridded predictions of population density. Formally, the population density in pixel  $i$  is estimated by

$$\rho_i^{RS} = \frac{w_i}{\sum_j w_j} P, \quad (24)$$

where  $w_i$  is the weight assigned to pixel  $i$  and  $P$  is the total population. Combining MP and RS data can produce population datasets with a high spatial and temporal resolution.

Douglass et al. [136] created high-resolution maps of population distribution by combining telecommunications data, satellite imagery and census data in Milan, Italy. They fitted population and call data by applying an elementary model that is similar to Eq. (23). They found that the total out-call volume has the strongest correlation (about 0.68) with the grid-level population. Further, they employed a random forest regression to predict population using features of land cover measures, call activity measures and their combinations. They found that building land cover and calls made out at 10am are the top-two predictors that are sufficient to provide accurate predictions. Lulli et al. [137] proposed a function to capture similarities between individual call profiles (ICPs). The similarity of ICPs is captured by combining the Euclidean similarity and the Jaccard similarity. Then, they built a clustering algorithm to provide clusters of individuals based on the similarity between ICPs. Using an automatic classifier to label the clusters, their method can estimate the number of residents, commuters and visitors in a given region. At the urban scale, Khodabandeh et al. [138] estimated population density by applying Eq. (22) and Eq. (23) based on the mobile network traffic metadata. Their method can estimate both static and dynamic populations across different cities.

Calling activities are powerful in mapping populations, however, it is usually not easy to obtain due to privacy concerns [139]. For example, some highly sensitive traits and attributes can be inferred from digital records of human behavior [140]. The increasingly available social media data presents alternative opportunities in estimating population distribution. Twitter has gained worldwide popularity, making the geotagged tweets show detailed depictions of human activity. Leetaru et al. [141] explored over 1.5 billion tweets posted by over 70 million users. They found a high correlation (0.79) between geotagged tweets and the NASA City Lights imagery. The most accurate feature is the self-reported user location field, exhibiting a correlation 0.72 with the geotagged baseline. Their work demonstrates the potential of geotagged tweets in world population mapping.

Very recently, volunteered geographic information (VGI) collected from the Internet (e.g., check-in data [142]) has been used to estimate population at a fine scale. Yao et al. [143] presented a framework to map population distribution at the building level by integrating national census data with two geospatial data sources. One is the points-of-interest (POIs) provided by Baidu Map Services, and the other is the real-time Tencent user densities (RTUD). They employed the random forest algorithm [144] to analyze the two geospatial datasets and downscale the street-level population distribution to the grid level. Then, they proposed an iterative gravity model that can efficiently estimates the population density in each building and study area. Their method achieves a high correlation to the official census data.

The WorldPop collection recently brings together publications describing detailed and open-access spatial demographic datasets built using transparent approaches [145]. For the Latin America and the Caribbean region, Sorichetta et al. [146] opened an archive of high-resolution gridded population datasets for 2010, 2015 and 2020 based on the most recent official population count data for 28 countries. Gaughan et al. [147] opened mainland China population maps for 1990, 2000 and 2010 after analyzing temporally-explicit census data using an ensemble prediction model. Lloyd et al. [148] described the datasets and production methodology for the 3 and 30 arc-second resolution global gridded population data. The basis of the archive contains four tiled raster datasets and other layers.

### 2.3.2. International migration

International migration is one of the major reasons of demographic, economic and political changes. Literature suggested some determinants of migration such as family and personal networks [149] and revealed the impact of immigrants on the host country's economy [150]. There are some bottlenecks in studying migration such as data availability, data quality, data collection rules, and inconsistencies in measurement. For example, a person may involve multiple migrations during a given year, but most systems considered the number of migrations instead of migrants, resulting in the overestimate of the amount of immigrants. Moreover, "migration" defined by different countries may differ substantially, which results in the inconsistencies among international data. In addition, which country is

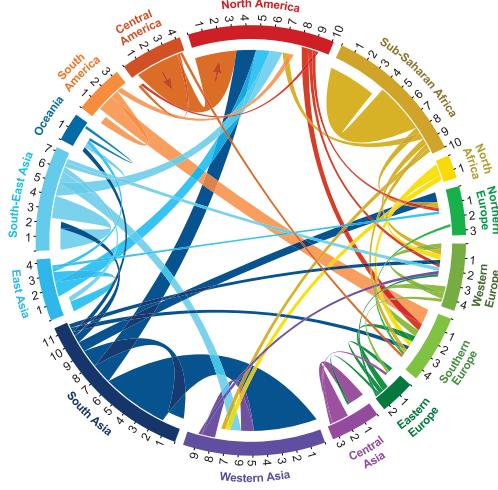


Figure 10: Circular plot of migration flows within and between world regions during 2005 to 2010. Tick marks show the number of migrants (inflows and outflows) in millions. Only flows containing at least 170,000 migrants are shown. Figure from [156].

reporting the data will lead to significant different patterns of migration [151]. Census and registered migration data are helpful for the estimation of international migrations. By combining census migration data and patient registration data, Raymer et al. [152] developed a log-linear model to estimate elderly migration flows in England and Wales. Their model extends the spatial interaction model (see Ref. [153] for details) by adding a third variable of interest, such as health status in migration data. Formally, the log-linear model with an offset is given by

$$\log(\lambda_{ijk}) = \log(\alpha_i) + \log(\beta_j) + \log(m_{ijk}), \quad (25)$$

where  $\lambda_{ijk}$  is the expected migration flow from origin  $i$  to destination  $j$  for level  $k$  of the third variable,  $\alpha_i$  and  $\beta_j$  are respectively related to the origin and destination's characteristics, and  $m_{ijk}$  is the auxiliary information on migration flows.

Cohen et al. [154] developed a generalized linear model (GLM) [155] to predict international migrants using only geographic and demographic variables. They found that the number of migrants per year depends on population of origin and its population density. De Beer et al. [151] presented a methodology to estimate total immigration and emigration numbers for 19 European countries. Abel and Sander [156] provided the spatial structure of international migration flows between 196 countries from 1990 to 2010 (see Figure 10). The bilateral migration flows are based on refugee statistics, population registers, and place-of-birth responses to census questions. They employed an iterative proportional fitting algorithm [157] to estimate the global migration flows. They found that the percentage of 5-year flows has been relatively stable at about 0.6% of world population since 1995. Moreover, African migrants move predominantly within the African continent, Asian and Latin American migration flows are spatially focused, and long-distance flows usually go to higher income level countries with negligible return flows.

The increasingly available geolocated digital records from intelligent devices and online platforms offer the opportunity to better quantify migration flows. Using a large sample of Yahoo! e-mail data, Zagheni and Weber [158] estimated the age and gender-specific migration rates. The locations of users are estimated by the country where their most messages were sent. The self-reported age and gender of users are then linked to their locations. They found that the estimated age profiles of migrants are consistent with the official data, and the mobility of females grows at a faster pace. Using the similar Yahoo! data of over 100 million users, State et al. [159] developed a statistical model to identify migrants and tourists. After generating a global mobility map, they found that the European Economic Area has high levels of pendularity, and pendular migrations are in closely located countries. State et al. [160] investigated international migration of professional workers by analyzing millions of geotagged career histories on LinkedIn. They found that the percentage of professional migrants to the US decreases from 2000 to 2012, while Asia has been a major professional migration destination during the past twelve years. Kikas et al. [161] extracted international migration from the Skype login events, showing that international migration can be estimated based on some social

network features such as the percentages of international calls. Barchiesi et al. [162] extracted the location of users from geotagged photographs on Flickr and inferred their trajectories. The estimated number of visitors to the UK correlates with the official estimates for 28 countries.

Twitter provides a rich source of geotagged data to estimate international migration. Based on about one billion tweets, Hawelka et al. [163] estimated the volume of international travelers according to the country of residence. They revealed spatially cohesive regions after analyzing the community structure of the Twitter-based international mobility network. By analyzing geotagged tweets produced by about 500,000 users, Zagheni et al. [164] evaluated recent trends of migrations in OECD countries. They applied a difference-in-differences approach [165] to reduce selection bias when inferring trends in out-migration rates. Their method can predict turning points in migration trends. Fagiolo and Mastrorillo [166] analyzed the topological structure of a international migration network (IMN) and its evolution from 1960 to 2000, where nodes are countries and links are the stock of migrants. They found that link weights follow a power-law distribution with a stable exponent at about 1.3. Moreover, IMN is highly clustered, disassortative, with a modular structure and of small-world property. In addition, most topological features of IMN can be explained by GLM, suggesting that socioeconomic, geographical and political factors are important in shaping the structure of migration networks.

International migration issues are prominent on economy and policy agendas. Fagiolo and Mastrorillo [167] studied how international migrations affect bilateral trade. They found that IMN and trade are strongly correlated with each other, and high centrality in IMN can increase the bilateral trade of countries. These results also indicate that the number of international immigrants can boost bilateral trade. Lee et al. [168] suggested the research themes to focus on the growth of migration flows driven by humanitarian crises and the connections between migration and inequality. The Global Migration Group [169] has provided guidance to support the collection, tabulation, analysis, dissemination and use of international migration data to monitor the implementation of the Sustainable Development Goals.

### 2.3.3. Culture evolution

Culture is the essential character of human society, and it serves as a driving force for human development. Quantifying cultural evolution is a challenging task due to the lack of suitable data. Recently, the development of information technologies has made large-scale data available for culture evolution studies [170] such as digitized books [171], baby names[172], languages [173], recipes [174], and biographies [175]. Moreover, human languages, as an important part of culture, have also been studied using novel data resources (e.g., social media data [176]) besides evolution models [177]. Here, we will introduce applications of new data sources and methods in quantifying culture evolution.

Part of the evolution of human society is recorded by books. By analyzing a corpus of 5 million Google digitized books, Michel et al. [171] observed cultural trends with over two billion culturomic trajectories. Focusing on linguistic and cultural phenomena reflected in the English language between 1800 and 2000, they provided insights about the size of English lexicon, collective memory, and evolution of grammar. In particular, the polarization of the states before the Civil War was revealed by the trajectories of using “the North”, “the South”, and finally “the enemy”. Zeng and Greenfield [178] analyzed massive culture-wide content using the Google Ngram Viewer. They found that cultural values shift along with specific ecological changes (urbanisation, wealth and formal education) in Chinese society. In particular, the frequencies of words related to adaptive individualistic values (indexed by words such as “choose”, “compete” and “get”) increases from 1970 to 2008. Bail [170] summarized text extraction methods to classify different types of culture and map cultural environments from text-based data. These new tools were further combined with conventional qualitative methods to track cultural element evolution. Figuring out sales trajectories of books will help understand the cultural evolution. Yucesoy et al. [179] revealed a universal sales pattern of bestsellers, and further proposed a model that can explain the time evolution of book sales.

Human behavior massively reflects cultural information. Schich et al. [180] reconstructed the aggregated mobility of over 150,000 intellectual individuals (see Figure 11) and then measured cultural interactions on a historical time scale. They developed quantitative methods to identify statistical regularities of individuals based on spatiotemporal birth and death information of notable individuals collected from Freebase.com (FB) and other sources (see Ref. [175] for a similar dataset of globally famous biographies). They found that the distribution of distances between the birth and death locations of notable individuals remains unchanged over eight centuries. By employing network tools and complexity theory, they further identified the characteristic statistical patterns. In particular, Europe can

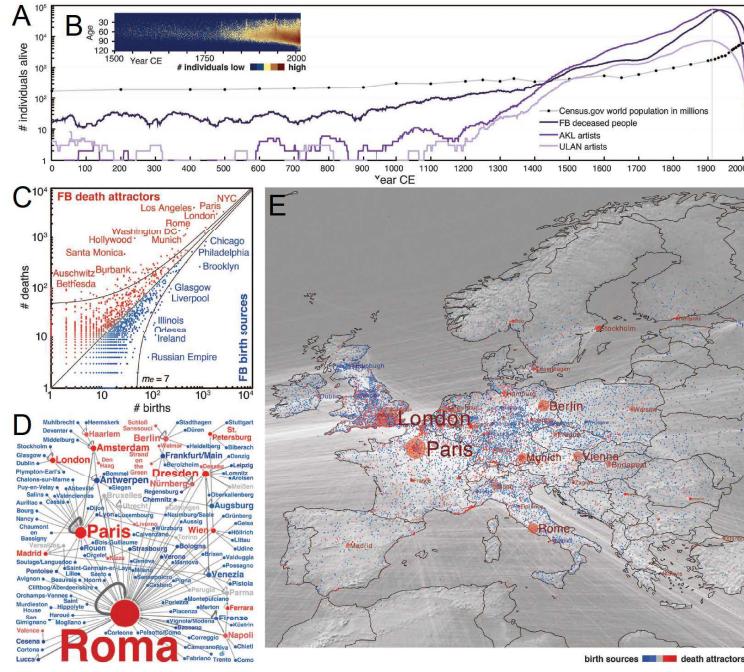


Figure 11: Interactions between culturally relevant locations over two millennia. (A) Notable individuals with birth and death locations. (B) Demographic life table for the Freebase.com (FB) dataset indicating death age frequency. (C) Birth-death scatter plot for locations in FB. (D) Illustration of birth-death flows of antiquarians in the 18th century. (E) Migration in Europe based on FB. Figure from [180].

be characterized by two different cultural regimes. One is winner-takes-all regime, where massive centralization is toward centers, and the other is fit-gets-richer regime, where many sub-centers compete in federal clusters. This work provides a macroscopic perspective of cultural history. Recently, Yang et al. [181] explored cultural mapping based on user behavioral data collected from location-based social networks. From check-ins messages and check-ins at a city's POIs, they extracted three key cultural features, namely, language usage, daily activity pattern, and intercity crowd mobility patterns. Then, they proposed a cultural clustering method to capture cultural features and generate cultural maps that match traditional survey-based ones.

Names can be used to study the underlying mechanism for cultural evolution. Hahn and Bentley [182] analyzed 1000 most commonly used baby names in the US in each decade of the twentieth century. They found that the frequency distribution of baby names obeys a power law for over 100 years, and the distribution can be explained by a simple process where names are randomly copied. Bentley et al. [183] explained the steady turnover of modern baby names using a random-copying model. Female names in each decade have a higher turnover rate than male names, implying more innovation in naming girls. The random-copying model can characterize collective copying behavior in culture evolution. Berger et al. [172] analyzed names given to babies born from 1882 to 2006 in the US. They found that the popularity of names is affected nonlinearly by the similar names that became popular recently. Xi et al. [184] found the sustained decline of inequality level among baby names with time. The reason behind this observation may be that people have more chances to know others' names, and new names need to be more distinctive and novel. Further, they proposed a stochastic model in which social influence and individual preference determine individual choice of names. Recently, Barucca et al. [185] analyzed the correlations of newborns' names in different states of the US from 1910 to 2012. They found a clear division of states into two homogeneous groups, where either group has similarity in their distributions of names. However, a transformation occurred at the end of the 20th century, where new clusters emerged in naming babies. Kim and Park [186] investigated the distribution of family names in Korea, finding that the growth rates of smaller family names are higher. Lee et al. [187] analyzed statistics of given names in Korea, Quebec, and the US. They found that the average popularities of given names show similar patterns of rise and fall at about one generation.

Language evolution is an important aspect of culture evolution. From a modeling perspective, Nowak et al. [188]

showed that some certain evolutionary dynamics can describe both the cultural evolution of language and the biological evolution of universal grammar. Abrams and Strogatz [189] developed a simple model of language competition that explains historical data on the decline of some endangered languages. They derived a linguistic parameter that quantifies the threat of language extinction for the model. From an empirical perspective, Lieberman et al. [190] quantified the evolving dynamics of language by analyzing the regularization of English verbs over the past 1,200 years. They explored how the rate of regularization depends on the frequency of word usage and found that the half-life of irregular verbs is proportional to the square root of their frequency. Based on the dataset of 107 million tweets, Eisenstein et al. [176] investigated the fundamental changes in the nature of written language. After employing a latent vector auto-regressive model to identify high-level patterns in the diffusion of linguistic change over the US, they found that language evolution in computer-mediated communication reproduces existing fault lines in spoken American English. Recently, Newberry et al. [191] quantified the strength of selection relative to stochastic drift in language evolution. After inferring selection towards the irregular forms of some past-tense verbs, they found that stochastic drift is stronger for rare words, suggesting that stochasticity plays an under-appreciated role in language evolution.

Vocabulary growth in natural languages follows scaling laws. For example, the character frequency distribution follows Zipf's law [192] in the relation  $Z(r) \approx r^\alpha$ , where  $r$  denotes the rank of a word by its frequency  $Z(r)$ , and  $\alpha$  is the Zipf's exponent. The number of distinct characters follows Heaps' law [193] as  $N(t) \approx t^\lambda$ , where  $N(t)$  denotes the number of distinct words when the text length is  $t$ , and  $\lambda \leq 1$  is the Heaps' exponent. Zipf's law and Heaps' law have been widely observed in Indo-European language family and keywords in journals [194]. Indeed, these two laws are mathematically related, say Zipf's law leads to Heaps' law [195]. After analyzing over 15 million words in books, Petersen et al. [196] found that only the more common words obey the classic Zipf's law, and the annual growth fluctuation of word usages decreases with the corpus size. Based on the Google Ngram database of books, Gerlach and Altmann [197] proposed a stochastic model for vocabulary growth that can generalize Zipf's and Heaps' law to two-scaling regimes. They found that the main historical change is the composition of specific words, where the list of core words is finite and decays exponentially in time with about 30 words per year for English. Pechenick et al. [198] analyzed the English fiction corpus and found that the Zipf's distribution has changed little from 1820 to 2000.

Some languages like Chinese, Japanese and Korean do not obey Zipf's law or Heaps' law. For these languages with very limited dictionary sizes (the number of characters is much smaller than the number of words), Lü et al. [199] found that  $Z(r)$  follows a power law with  $\alpha \approx 1$ , and  $N(t)$  grows with the text length in three stages:  $N(t)$  grows linearly at the beginning, then turns to a logarithmic form, and saturates in the end. After analyzing four Chinese texts, Deng et al. [200] found that Zipf's law perfectly holds for sufficiently short texts of Chinese characters. However, rank-frequency relations display a two-layer structure for long texts, with a Zipfian power-law regime for high-frequent characters in the first layer and an exponential-like regime for less-frequent characters in the second layer. Yan and Minnhagen [201] proposed a neutral model to predict character frequency distributions in *Chinese characters*, where the maximum entropy prediction is used to describe a text written in Chinese. They demonstrated that the same Chinese texts written in *words* and *Chinese characters* are both well predicted by their three characteristic values (the total number of words, the number of distinct words, and the number of repetitions of the most common word). Yan et al. [202] further built a node-weighted network of Chinese characters, in which the weights of nodes are the frequencies of character usages, and the directed links correspond to the relations of direct components of characters. They developed a distributed node weight (DNW) strategy for learning Chinese characters and analyzed learning strategies using the dynamical processes. Results showed that the DNW strategy can significantly improve the efficiency of learning major Chinese textbooks.

Language can be used to reveal linguistic and cultural borders. Bryden et al. [203] explored the interlink between language and social network structure based on Twitter data. They found that the hierarchy of communities on social networks can be characterized by their most significantly used words. The community of a user can be predicted by the used words in tweets. Based on co-editing activities of Wikipedia, Samoilenco et al. [204] studied the linguistic neighbourhoods between language communities. They found that similar interests of Wikipedia editors between cultural communities can be explained by bilingualism, linguistic similarity of languages, and shared religion. Further, they proposed a method that can extract cultural borders from the co-editing activities. Mocanu et al. [205] studied worldwide linguistic indicators and trends (see Figure 12) by analyzing a large-scale dataset of geotagged tweets. They found that Twitter penetration is highly heterogeneous and it is strongly correlated with GDP. Moreover, tweets can be used to study linguistic homogeneity at the country level, map language distributions in regions, and identify

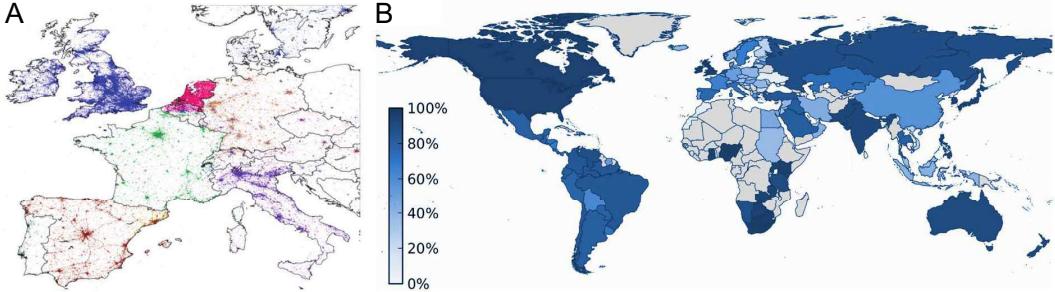


Figure 12: Geographic distribution of languages based on Twitter data. (A) Raw Twitter signal. Each color represents a language. Densely populated areas are easily identified, showing that languages are separated among European countries. (B) Dominant language usage. The color indicates the fraction of users adopting the official language in tweets. Figure from [205].

linguistically specific communities in urban areas.

Language can also reflect the ability of cultural influence. Ronen et al. [173] proposed a quantitative measure for a language’s global influence based on the structure of three global language networks (GLNs). The GLNs are constructed by identifying significant links between languages with respect to the population of speakers expressed in three datasets. Formally, the correlation  $\phi_{ij}$  between languages  $i$  and  $j$  is given by

$$\phi_{ij} = \frac{M_{ij}N - M_iM_j}{\sqrt{M_iM_j(N - M_i)(N - M_j)}}, \quad (26)$$

where  $M_{ij}$  is the number of multilingual users (or book translations) between languages  $i$  and  $j$ ,  $M_i = \sum_j M_{ij}$ , and  $N$  is the total number of users (or book translations). The statistical significance of the correlation is given by the  $t$  statistic,

$$t_{ij} = \frac{\phi_{ij} \sqrt{D - 2}}{\sqrt{1 - \phi_{ij}^2}}, \quad (27)$$

where  $D - 2$  is the degree of freedom and  $D = \max(M_i, M_j)$ . Empirical results show that the position of a language in the GLNs contributes to the visibility of its speakers and the global popularity of the cultural content they produce. Gonçalves et al. [206] explored a large corpus of geotagged tweets and the Google Books datasets corresponding to books published in the US and the UK. After studying how the world-wide varieties of written English are evolving, they found that the past two centuries have clearly resulted in a clear shift in vocabulary and spelling conventions from British to American. The result suggests the capacity to culturally influence the rest of the world gradually shifts from the UK to the US.

Food is an integral part of cultures. Counihan and van Esterik [207] analyzed food-related activities and presented a crosscultural study of personal identities and social groups. They introduced empirical and theoretical tools to understand food systems at multiple levels. Data of cuisines have been used to study food culture. Ahn et al. [208] explored cultural diversity by analyzing the variety of regional cuisines. After introducing a flavor network to capture the ingredient combinations in recipes, they found that Western cuisines tend to use compound sharing ingredients, supporting the food pairing hypothesis that ingredients having similar flavor compounds may taste well together [209]. By contrast, East Asian cuisines show a tendency to avoid food pairing. Zhu et al. [174] explored the similarity of regional cuisines in China based on online recipes. They found that geographical proximity plays a more crucial role than climate proximity in determining regional cuisine similarity. Further, they proposed an evolution model of Chinese cuisines that achieves the similar tendency as the real dataset. Their work extends our understanding of the evolution of Chinese regional cuisines and cultures.

Food preference can reflect cultural diversity and cross-cultural relations. Based on a server log data from a large recipe platform, Wagner et al. [210] explored the evolution of food preferences. They found that ingredients partly drive recipe preferences, and ingredient preference distributions have less regional differences than recipe preference distributions. Moreover, weekday preferences differ from weekend preferences. Abbar et al. [211] studied US-wide dietary choices by analyzing dining experiences tweeted by 0.21 million users. They found that the caloric

values of tweeted foods have a high correlation (0.77) with the state-wide obesity rates. Moreover, users in higher-educated areas tweeted about food with less caloric. Based on twitted food names and demographic variables, they built a model that can well predict county-wide obesity and diabetes statistics. Laufer et al. [212] explored the cross-cultural relations based on 31 European food cultures recorded by Wikipedia. They mined cultural relations through the collective description and popularity of culinary practices within and across different Wikipedia language communities. They found that shared internal states (e.g., beliefs and values) are positively correlated with shared culinary practices, and neighbouring countries tend to have similar cultural practices.

### 3. Regional socioeconomic status and urban perception

#### 3.1. Economic activity and socioeconomic status

High-resolution data and improved methods allow us to reveal economic activity and socioeconomic status in subnational, regional and urban scales. In this section, we will introduce recent works on predicting regional economic activity from nighttime lights (NTLs), mapping slums from very high resolution (VHR) imagery, inferring regional socioeconomic status from mobile phone (MP) data, and quantifying regional economic development based on social media (SM) data.

##### 3.1.1. Nighttime lights reflect economic activity

Nighttime lights (NTLs) data have been widely used to infer socioeconomic status and predict income per capita at the regional and urban scales. Sutton et al. [213] proposed a regression model to estimate subnational GDP of India, China, Turkey, and the US. They estimated urban population using a log-log relationship between population and the areal extent of lit areas in NTLs imagery derived from the Defense Meteorological Satellite Program-Operational Line Scan System (DMSP-OLS) [214]. They predicted GDP for the subnational administrative units by adding the estimated urban population into the regression model, with an assumption that urban population is the most critical factor for economic activity. They found that spatial disaggregation of estimates dramatically improved the aggregated national estimates of GDP based on NTLs.

Regional and urban socioeconomic status can be inferred from the changes in electric power consumption patterns reflected by NTLs that are derived from the DMSP-OLS data. Chand et al. [215] studied the socioeconomic development of states and cities in India by looking at the spatial and temporal changes in electric power consumption. They found that the number of NTLs overall increases up to 26% in all states from 1993 to 2002, but there is a decline in some states. The increase in population correlates with both the increase in NTLs ( $R^2 = 0.59$ ) and the electric power consumption ( $R^2 = 0.56$ ). For the Republic of Kazakhstan, Propastin and Kappas [216] leveraged NTLs to monitor socioeconomic indicators (e.g., population, electricity consumption and GDP) at different spatial resolutions. Linear regression models were used to estimate population and electricity consumption at the settlement level. They revealed a strong correlation between NTLs and GDP. In particular, the regression model can explain 76% of the spatial variability in GDP among 17 provinces and 94% of the inter-annual variation in total GDP of Kazakhstan during 1994-1999.

Luminosity from NTLs satellite imagery has been used as a proxy for economic statistics. Chen and Nordhaus [217] studied how much luminosity can contribute to the construction of GDP measures. They proposed an analytic method to quantify the relationship between luminosity from the DMSP-OLS data and GDP of North America. They found that luminosity is likely to add values as a proxy to estimate economic output for countries and regions. Due to a high measurement error of luminosity, however, the added values are limited for countries and regions with poor quality data. For more than 200 cities in China, Ma et al. [218] comparatively used three regression models to study the responses of stable NTLs from the DMSP-OLS data to changes in urbanization variables (e.g., population, GDP, electric power consumption, and built-up area). They found that NTLs can help estimate urbanization dynamics.

Mellander, et al. [219] studied the relationship between NTLs and economic activity at a fine level. They used a geo-coded socioeconomic dataset consisting of spatially matched population and establishment counts in Sweden. After matching the dataset with light emissions, they used correlation analysis and geographically weighted regressions (GWR) to examine the relationship. The GWR model is given by

$$y_i = \beta_0(i) + \beta_1(i)x_{1i} + \beta_2(i)x_{2i} + \dots + \beta_n(i)x_{ni} + \varepsilon_i, \quad (28)$$

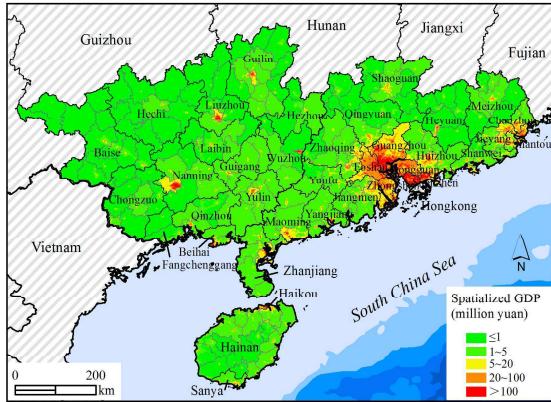


Figure 13: The pixel-level ( $500\text{m} \times 500\text{m}$ ) GDP map for South China in 2014. The GDP map was produced by using the corrected NPP-VIIRS data and the regression model. Figure from [221].

where  $i$  denotes location,  $y$  is the dependent variable of NTLs,  $x$  is a population-based or industry-based socioeconomic variable, and  $\varepsilon$  is the error term. They found that NTLs is a good proxy for population density and a weaker proxy for economic activity at the micro-level. The link between NTLs and economic activity is slightly overestimated for large urban areas but underestimated for rural areas. Moreover, economic activity has a stronger correlation with radiance light than saturated light.

NTLs satellite imagery has been used to study the economic differences and GDP spatialization at the regional levels in China. For example, Zhao et al. [220] examined the relations between China's environmental change and economic development from 1996 to 2000. A proxy evaluator of ecosystems is net primary production (NPP), which is the amount of carbon and energy that enters ecosystems. A proxy evaluator of economic development is GDP, which is estimated based on NTLs. They found that changing patterns of NPP and GDP vary by regions, and the relations are greatly affected by spatial locations. Later, Zhao et al. [221] studied economic differences among diverse geomorphological types in South China based on NTLs from the Visible Infrared Imaging Radiometer Suite (VIIRS) [222]. They found that the total NTLs exhibit a high correlation with both prefecture GDP ( $R^2 = 0.8935$ ) and county GDP ( $R^2 = 0.9243$ ), suggesting the capability of NTLs in estimating economic development. Further, they proposed a GDP spatialization model and produced a pixel-level GDP map for South China (see Figure 13). Meanwhile, Dai et al. [223] explored the suitability of using NTLs and regression methods to estimate GDP at the provincial and city levels in China. Empirical analysis suggested that the VIIRS data outperforms the DMSP-OLS data, and the polynomial model outperforms the linear regression model.

### 3.1.2. Very high resolution imagery maps poverty

Indicators derived from both nighttime lights (NTLs) and very high resolution (VHR) imagery have been used to map poverty at fine scales. By analyzing stable NTLs, Weng et al. [224] estimated poverty for 31 provinces in China. Specifically, they used the principal component analysis (PCA) to develop a so-called Integrated Poverty Index (IPI) based on 17 socioeconomic indexes and computed the average light index (ALI) for each region based on the total number of lit pixels. They found a high correlation ( $r = 0.85$ ) between IPI and ALI for the 31 provinces, suggesting the capacity of NTLs imagery to analyze poverty at the regional level. Using a supervised learning approach, Engstrom et al. [225] linked features derived from VHR satellite images to survey-based poverty rates at the local level in Sri Lanka. They found that satellite-based features are highly predictive to poverty. Moreover, measures of built-up area and building density are strongly correlated with welfare in both urban and rural areas.

Slums are common in low and middle income countries with poor quality of basic services (e.g., water supply, electricity and sanitation). Detecting and monitoring slum areas is valuable for implementing policies to improve living conditions. However, mapping the spatial distribution of urban slums is a challenging problem due to the reasons that there is no universal model of slums and slum conditions can take various forms. The meeting in 2008 [226] brought together the methodological expertise on slum monitoring and reviewed methods for slum identification based on VHR imagery. Using object-oriented classification of VHR images, Rhinane et al. [227] developed a novel

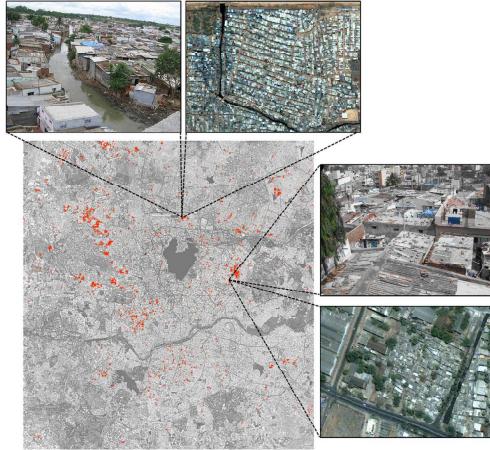


Figure 14: Slum map in Hyderabad, India. The slum locations (red areas) are identified by the lacunarity-based slum detection algorithm. Two different subsets of the original satellite image together with georeferenced photographs are shown as ground truth: the Rasolpoora slum in the northern and the Nagamiah Kunta slum in the eastern. Figure from [231].

approach to detect slums in Casablanca, which integrates spectral, spatial and contextual information to map the urban land, achieving a high accuracy 0.85 in extracting slum areas.

VHR images have been increasingly used to inventory the location and physical composition of slums. Shekhar [228] applied the object oriented analysis [229] to VHR images to detect slums in Pune, India. First, they generated segments by automatically dividing images into coherent objects. Then, they used the feature extraction method to identify the characteristic features for object-classes. Finally, they used contextual information to separate slum objects from non-slum built-up objects. Their approach exhibits the overall accuracy 87% in the classification of slums. Kohli et al. [230] developed an ontological framework to conceptualize slums based on input from 50 domain-experts covering 16 different countries. They identified the morphology of built environment at the environs, settlement and object levels. By including all potentially relevant indicators, their ontological framework provides a comprehensive basis for image-based classification of slums.

Recent literature have applied advanced image processing techniques to map slums from VHR images with minimal operator intervention. Kit et al. [231] developed the concept of lacunarity to identify slums in Hyderabad, India. First, they produced high resolution binary image using two binarization methods, the principal component analysis (PCA)-based method and the line detection-based method [232]. Then, they calculated lacunarity based on the binary image following Malhi and Román-Cuesta [233]. Formally, the lacunarity  $\Lambda$  of a subset  $P$  of the original binary image is given by

$$\Lambda = \frac{\sigma_r}{\bar{x}_r^2} + 1, \quad (29)$$

where  $\sigma_r$  is the variance and  $\bar{x}_r$  is the arithmetic mean of the number of filled pixels within all  $r$ -sized unique square subsets of the larger subset  $P$  (see Ref. [231] for details). The line detection algorithm performs better than the PCA-based method in providing suitable binary datasets for lacunarity analysis. The best method can reach an accuracy 0.8333 in slum identification when  $\Lambda \in [1.10, 1.15]$ . Figure 14 shows the slum map of Hyderabad generated by the lacunarity-based slum detection algorithm.

Kit et al. [234] soon improved the lacunarity-based slum detection algorithm by combining two advanced image analysis methods (the Canny edge detection [235] and the line-segment-detection (LSD) straight line detection [236]) to reduce errors in slum identification. Their method identifies the plausibly and spatially explicit slum locations, which can be verified by a series of ground truthing visits. In particular, such method can capture the changing patterns of slum areas from 2003 to 2010 in Hyderabad, India. Gruebner et al. [237] mapped urban slums in Dhaka, Bangladesh, from the visual interpretation of Quickbird data from 2006 to 2010. To avoid small and isolated slums, they filtered the 2006 slums in GIS and defined the changes of 2010 slums over the 2006's polygons to retain border consistency. Accordingly, they produced a slum distribution dataset for the Dhaka metropolitan area.

Engstrom et al. [238] mapped slum areas in Accra, Ghana, by utilizing features extracted from the VHR Quickbird images acquired in 2002. They demonstrated that the satellite image-derived slum areas exhibits an overall accuracy of 94.3% when comparing to the field-based slum map from the UN Habitat/Accra Metropolitan Assembly (UNAMA). However, the accuracy drops when comparing to two census derived slum maps. Moreover, they found a moderate correlation ( $r = 0.67$ ) between satellite image-derived classification of slums and the census derived slum index, and the correlation increases ( $r = 0.88$ ) after taking into account population density. Kohli et al. [239] studied the spatial uncertainties related to slum delineations, which are observed from VHR images in Ahmedabad (India), Nairobi (Kenya) and Cape Town (South Africa). They found that the slum identification and delineation for the three contexts are significantly different, suggesting the existential and extensional uncertainty of slums.

VHR imagery allows the monitoring of slums and the analysis of deprived areas. Kuffer et al. [240] utilized the gray-level co-occurrence matrix (GLCM) variance to distinguish slums areas in VHR imagery. They showed that the GLCM variance combined with the normalized difference vegetation index (NDVI) can separate slum areas with an overall accuracy 87%, 88% and 84% for Mumbai (India), Ahmedabad (India) and Kigali (Rwanda), respectively. The overall accuracy can be increased to 90% by adding spectral information to the GLCM within a random forest classifier [144]. Wurm et al. [241] explored the capabilities of X-band Synthetic Aperture Radar (SAR) data to estimate the extent of poverty in slum areas using the Kennaugh element framework in image preprocessing [242]. Employing a random forest classifier, they tested different spatial image features at various window sizes to map slums. Results show that GLCM performs very well on slum mapping as it addresses a large spatial neighborhood of the pixels.

Recently, Kuffer et al. [243] provided a literature review of slum mapping regarding four dimensions: contextual factors, physical slum characteristics, data and requirements, and slum extraction methods. They argued that the diversity and dynamics of slums have not been well captured due to the complex and diverse morphology of slums. Thereby, a more systematic exploration of physical slum characteristics is required (see Ref. [243] for details). They demonstrated that texture-based methods show good robustness, while machine-learning algorithms exhibit the highest reported accuracy. Mahabir et al. [244] suggested to develop a more comprehensive framework by considering emerging sources of geospatial data (e.g., social media) and combining multiple emerging approaches in technology (e.g., geosensor networks).

### 3.1.3. Mobile phones track socioeconomic levels

Scientists have explored the relations between social structure and economic development. Woolcock [245] provided a brief intellectual history of social capital and economic development. Adler and Kwon [246] synthesized studies on social capital undertaken in various disciplines and developed a common conceptual framework. Granovetter [247] suggested several underlying mechanisms on how social structure affects economic outcomes, for example, social networks influence the flow and the quality of information, and social networks are an important source of reward and punishment. Recently, empirical works have demonstrated that social network analysis of large-scale mobile phone (MP) data can be applied to monitor socioeconomic development.

Based on MP data and socioeconomic metric from national census, Eagle et al. [248] investigated the relation between the structure of communication network and economic development at the population level in the UK (see Figure 15). The socioeconomic metric is the 2004 UK government's index of multiple deprivation (IMD), which is a composite measure of relative prosperity of communities. They calculated the socioeconomic profile of a region by aggregating the population-weighted average of the IMD for each telephone exchange area. The communication network data covers over 90% of MP users during August 2005, based on which they calculated two diversity metrics of communication ties. The social diversity  $D_{\text{social}}(i)$  is defined as the Shannon entropy  $H(i)$  associated with individual  $i$ 's communication behavior normalized by its number of contacts  $k$ . Formally,

$$D_{\text{social}}(i) = \frac{H(i)}{\log(k)} = \frac{-\sum_{j=1}^k p_{ij} \log(p_{ij})}{\log(k)}, \quad (30)$$

where  $p_{ij}$  is the proportion of individual  $i$ 's call volume that involves individual  $j$ . A regions's social diversity is then calculated by averaging the social diversities of individuals in that region. The spatial diversity  $D_{\text{spatial}}(i)$  is defined by

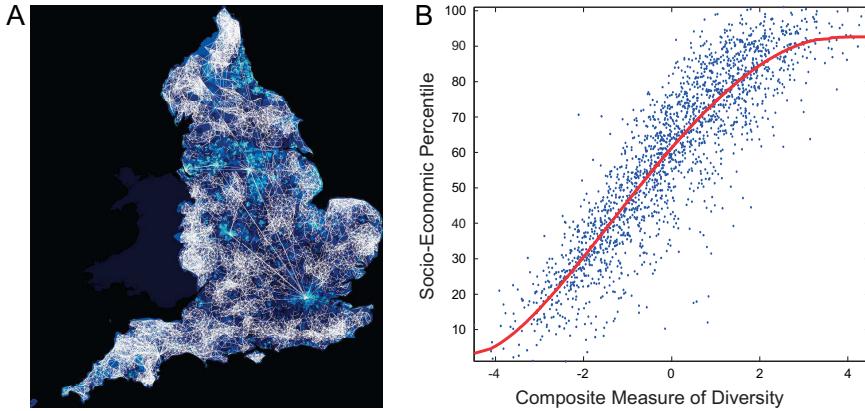


Figure 15: Regional communication diversity and socioeconomic ranking for the UK. (A) Communication networks of regions and regional socioeconomic rank based on IMD. Saturation and width of links correspond to the volume of communications. High rank to low rank of the IMD is represented by light blue to dark blue. (B) The relation between social network diversity and socioeconomic rank. The network diversity was a composite of Shannon entropy and Burt's measure of structural holes. The fractional polynomial fit to the data is shown in red. Figure from [248].

replacing call volume with geographic distance. Formally,

$$D_{\text{spatial}}(i) = \frac{-\sum_{a=1}^A p_{ia} \log(p_{ia})}{\log(A)}, \quad (31)$$

where  $A$  is the number of exchange areas, and  $p_{ia}$  is the proportion of time that individual  $i$  spent on communicating with area  $a$ . They found that the IMD socioeconomic rank is strongly correlated with both the social diversity  $D_{\text{social}}$  ( $r = 0.73$ ) and the spatial diversity  $D_{\text{spatial}}$  ( $r = 0.58$ ). The strong correlation ( $r = 0.72$ ) persists if using Burt's measure of "structural holes" (see Ref. [249] for details). Moreover, a composite diversity measure can exhibit even stronger correlation ( $r = 0.78$ ) with socioeconomic status (see Figure 15B). This work takes a significant step towards inferring regional socioeconomic status from MP data.

The ubiquitous adoption of MPs in emerging economies provides a new way to track socioeconomic status. Based on CDRs of 0.22 million MP users in an advanced economy and 0.19 million MP users in a developing economy, Rubio et al. [250] studied human mobility patterns in regions of different socioeconomic levels. They found that individuals in the developing economy have smaller average traveled distance, their social networks have smaller geographical sparsity, and these patterns have no significant changes from workweeks to weekends. Later, Frias-Martinez and Virseda [251] explored the relations between behavioral features extracted from large-scale CDRs and socioeconomic indices from country-wide census data in a Latin American country. They found that socioeconomic levels are strongly correlated with expenses, reciprocity of communications, physical distance with contacts, mobility patterns, and some others. Moreover, a multivariate linear regression including MP usage variables can accurately predict census-based variables such as the socioeconomic level ( $R^2 = 0.83$ ). These results suggest MP-derived human mobility patterns can be used to predict socioeconomic indices at fine scales.

A body of literature have leveraged CDRs to estimate regional socioeconomic status. A widely used CDRs data contain 2.5 billion calls and SMS exchanges from anonymous customers in Côte d'Ivoire. Smith-Clarke et al. [252] estimated socioeconomic levels of regions in Côte d'Ivoire. They derived some socioeconomic-related features from the communication flows, including activity, gravity residual, network advantage, and introversion. They found that regions with higher call volumes from other regions are more likely to have a higher socioeconomic level. Further, they proposed a simple linear model that estimates socioeconomic status for 255 sub-prefectures in Côte d'Ivoire. Šćepanović et al. [253] extracted different spatial-temporal mobility patterns from the same CDRs in Côte d'Ivoire and used them to predict socioeconomic indices. They showed that the spatial-variance of calling frequency can identify electricity lacking rural and regions, the spatial-variance of the probability density functions of the radius of gyration (see Ref. [55] for the definition) can identify a region's wealth, and the number of a region's migration workers is negatively correlated ( $r = -0.7681$ ) with the multidimensional poverty index (MPI).

Recently, MP data have been combined with other data sources to study socioeconomic stratification. Leo et

al. [254] analyzed a coupled datasets of MP communications and bank transactions for over one million people in Mexico. They constructed a social network based on call/SMS interactions and estimated economic indicators based on bank transactions. After calculating the cumulative distributions of individual average monthly purchase (AMP) and debt (AMD), they found that both wealth and debt are unevenly distributed among people. Further, they studied the social stratification by categorizing users into nine socioeconomic classes using the cumulative AMP function. They observed that people are more densely connected to others of their own class. To quantify this observation, they calculated the “rich-club” coefficient [64],

$$\rho(P_>) = \frac{\phi(P_>)}{\langle\phi_m\rangle(P_>)}, \quad (32)$$

where  $\phi(P_>) = 2L_{P_>}/N_{P_>} (N_{P_>} - 1)$  and  $\langle\phi_m\rangle(P_>)$  is the average density. Here,  $L_{P_>}$  and  $N_{P_>}$  are respectively the number of links and nodes remaining in the communication network after removing nodes with their AMP value  $P_u$  smaller than a given threshold  $P_>$  (see Ref. [254] for details). They found that the rich-club coefficient grows rapidly with  $P_>$ , suggesting an assortative socioeconomic correlation.

Moreover, Leo et al. [254] studied the spatio-socioeconomic correlations by calculating the average geodesic distance between any pairs of socioeconomic classes,

$$\langle d_{\text{geo}}(s_i, s_j) \rangle = \frac{1}{|E(s_i, s_j)|} \sum_{\substack{(u,v) \in E \\ u \in s_i, v \in s_j}} d_{\text{geo}}^{\text{zip}}(u, v), \quad (33)$$

where  $|E(s_i, s_j)|$  is the number of links between nodes in classes  $s_i$  and  $s_j$ , and  $d_{\text{geo}}^{\text{zip}}(u, v)$  is the geodesic distance between zip locations of individuals  $u$  and  $v$ . They found that the distance  $\langle d_{\text{geo}}(s_i, s_j) \rangle$  is always minimal between individuals of the same class, suggesting that individuals from the same socioeconomic class live relatively the closest. In addition, there is a positive correlation between individuals’ socioeconomic levels and their typical commuting distances. After further exploring the same coupled dataset, Leo et al. [255] found a strong correlation between identified socioeconomic classes and typical consumption patterns.

### 3.1.4. Social media reveals socioeconomic status

Social media (SM) data have many appealing advantages including low acquisition cost, wide geographical coverage and real-time update, which enable the feasibility to estimate socioeconomic status at regional and urban scales. For example, Twitter provides a huge number of tweets with user locations being directly tagged or can be mined out from content information. Cheng et al. [256] proposed a probabilistic framework to automatically identify words related to locations in tweets and then infer a Twitter user’s location at the city level from the content. They showed that about one hundred tweets are enough for their method to infer a user’s location. This method can place on average 51% users within 100 miles of their actual locations.

The contents of SM posts have been used to track socioeconomic well-beings. Quercia et al. [257] studied the relations between sentiment expressed in tweets and census-based socioeconomic well-being of communities in London. Specifically, they calculated the word count sentiment score [258] by counting the number of positive and negative words. Formally,

$$S_i^{\text{WC}} = \frac{p_i - \mu_p}{\sigma_p} - \frac{n_i - \mu_n}{\sigma_n}, \quad (34)$$

where  $p_i$  ( $n_i$ ) is the fraction of positive (negative) words for user  $i$ ,  $\mu_p$  ( $\mu_n$ ) is the mean of  $p$  ( $n$ ) across all users, and  $\sigma_p$  ( $\sigma_n$ ) is the corresponding standard deviation. Then, the gross community happiness (GCH) of a community is calculated by averaging the sentiment scores of users in that community. The GCH is highly correlated with the community’s socioeconomic well-being, suggesting the effectiveness of using tweets to track community well-being.

Mahmud et al. [259] inferred home locations of Twitter users at different granularities using an algorithm that ensembles statistical and heuristic classifiers [260]. The algorithm achieves a higher performance in predicting Twitter users’ locations compared with the state-of-the-art algorithms. Hasan et al. [261] analyzed human activity patterns based on tweets with location information. By finding the distributions of different activity categories over a city geography, they characterized aggregate activity patterns and determined the purpose-specific activity distribution maps. Moreover, the timing distribution of visiting different places depends on activity category. Hasan and Ukkusuri [262] further proposed a data-driven modeling approach based on topic models [263] to infer urban activity pattern

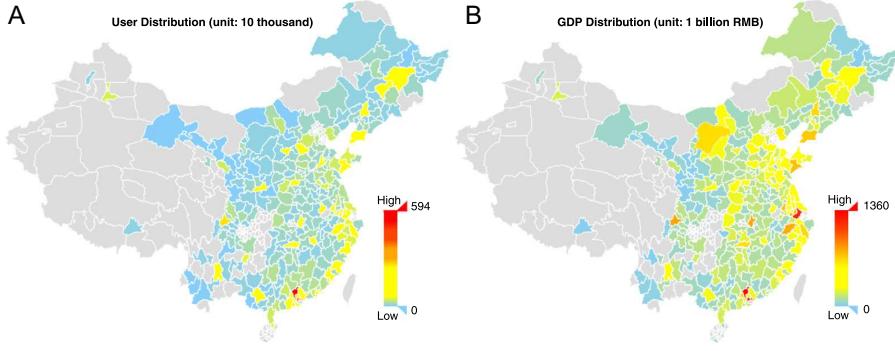


Figure 16: The spatial distributions of (A) the number of registered users in Weibo and (B) the values of GDP in the 282 prefecture-level cities of China in 2012. Figure from [266].

from geotagged tweets. Results demonstrated that their model can extract user-specific activity patterns and predict missing activities.

Using Twitter data generated during weekdays in Inner London, Lansley and Longley [264] applied an unsupervised learning algorithm to classify geo-tagged tweets into 20 distinctive and interpretive topic groupings. They found that users' socioeconomic characteristics can be inferred from their behaviours on Twitter. In particular, users whose neighbourhoods are of higher socioeconomic levels tend to tweet optimistically and discuss business, networking and leisure. Huang and Wong [265] explored to what extent Twitter data can be used to support the activity pattern analysis of users with different socioeconomic status. Activity patterns of Twitter users in Washington, D.C. were analyzed, and their socioeconomic levels were inferred by incorporating census data. Results showed that socioeconomic status remarkably affects users' activity patterns. Moreover, the urban spatial structure is a key factor that affects the variation in activity patterns among users from different communities. In particular, the mid-income group other than the most affluent group may have the shortest travel. Moreover, affluent residents are more internationally oriented than mid-income and poor residents.

Liu et al. [266] collected the registered location information of nearly 200 million Weibo users from 2009 to 2012 and explored the relationship between online activities and socioeconomic indices. Specifically, the online activity is estimated by the number of registered users (UN), and the socioeconomic indices are resident population (RP), GDP and GDP per capita. Figure 16 presents the spatial distributions of registered Weibo users (left) and the values of GDP (right). After calculating two correlation coefficients (Pearson coefficient  $r$  [267] and Spearman's rank coefficient  $\rho$  [268]), they showed that UN is strongly correlated with socioeconomic indices. For example, the strengths of correlation between UN and GDP are  $r = 0.88$  and  $\rho = 0.90$ . These results demonstrate that socioeconomic status can be inferred from online social activity at the city-level. Of particular significance, they further proposed a method to detect a few abnormal cities, whose GDP is much higher than others with the same number of registered users. These GDP winners have less-diverse economic structure and highly dependent on some specific resources. In fact, these cities' economics experienced a huge loss after 2013 due to the market price fluctuation of non-renewable energy resources and rare earths.

The structure of location-based social networks (LBSN) has been linked to socioeconomic development. Wang et al. [269] estimated regional economic status based on the structures of information flow and talent mobility networks (see Figure 17). Specifically, the online information flow network is built on the following relations among about 433 million Weibo users (see also Ref. [266]), and the offline talent mobility network is built on the resumes of about 142 thousand anonymized Chinese job seekers with higher education (see Ref. [270] for details). They calculated ten network structural features such as spatial and topological diversities and then linked them to regional economic indices. They found that structural features of both networks are relevant to economic status, while the talent mobility network exhibits a stronger predictive power for regional GDP. Further, they constructed a composite index of structural features, which can explain up to about 84% of the variance in regional GDP.

Based on data from the SM platform Gowalla [271] with friendship information and geo-locations, Holzbauer et al. [272] studied the relations between regional economic status and quantitative measures of social ties in the US

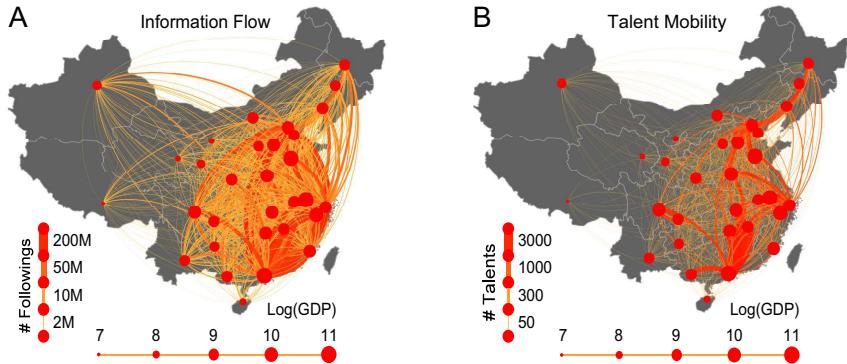


Figure 17: Networks of online information flow and offline talent mobility. Nodes represent provinces, where the size of node corresponds to the province's GDP in natural logarithmic form in 2016, and the layout of node corresponds to the geographical location of the province's capital city. (A) The online information flow network, where the link weight corresponds to the number of followings on Weibo. (B) The offline talent mobility network, where the link weight corresponds to the number of moved talents as recorded by their resumes. Figure from [269].

during 2009-2012. They found that cross-state long ties are strongly correlated with three economic measurements, namely, GDP ( $r = 0.921$ ), the number of patents ( $r = 0.788$ ), and the number of startups ( $r = 0.892$ ), while short ties are much less predictive. This finding highlights the role of long ties in supporting regional innovation and economic development. Recently, Norbutas and Corten [273] explored the relations between network structure and economic prosperity of 438 municipalities in Netherlands by analyzing data of over 10 million users on the Dutch online social network Hyves. They found that network diversity in terms of geographical distance [274] other than contacts' topological diversity [248] exhibits a positive correlation with economic prosperity, while network density at the community level and network modularity [275, 276] are negative predictors of economic status.

SM data have also been used to measure socioeconomic deprivation of regions (e.g., low level of economic status and lack of education) and quantify landscape values (e.g., the values shaped by the recreational and cultural services and benefits provided by landscapes). Venerandi et al. [277] proposed a method to automatically mine deprivation from two datasets of urban elements in physical environment at a fine level in UK. The two datasets are respectively collected from Foursquare, a mobile social-networking application with check-ins, and OpenStreetMap, an openly global accessible map with geographical positions, names and categories. They defined the offering advantage to identify distinctive urban elements of each neighborhood (see Ref. [277] for details) and built accurate classifiers of urban deprivation that can be verified by the census-based IMD. Later, van Zanten et al. [278] analyzed data from three online SM platforms (Panoramio, Flickr and Instagram). They found that data from these three platforms reveal similar patterns of landscape values. In particular, a significant portion of observed variation across different platforms can be explained by variables describing accessibility, population density, income, mountainous terrain, proximity to water, and so on.

### 3.2. Industrial structure and development path

Data from many new sources have been used to quantify economic structure and analyze industrial diversification, including large-scale social media data, labor market data, trade data, publicly listed firm data, and so on. In this subsection, we will briefly introduce the quantification of regional industrial structure, the role of relatedness on economic diversification, the collective learning effects, and the strategies for regional economic development.

#### 3.2.1. Economic structure and relatedness

Economic development is not only a process of continuously improving the production of the same goods and the occupation of the same industries [16], but also one that requires structural transformation toward new economic activities associated with higher levels of productivity [76, 77]. This implies that economic development and industrial diversification is a path-dependent process where structural transformation plays an important role. Revealing industrial structure of regions and quantifying relatedness of industries are critical for understanding development paths of regions and evolution patterns of regional economic diversification.

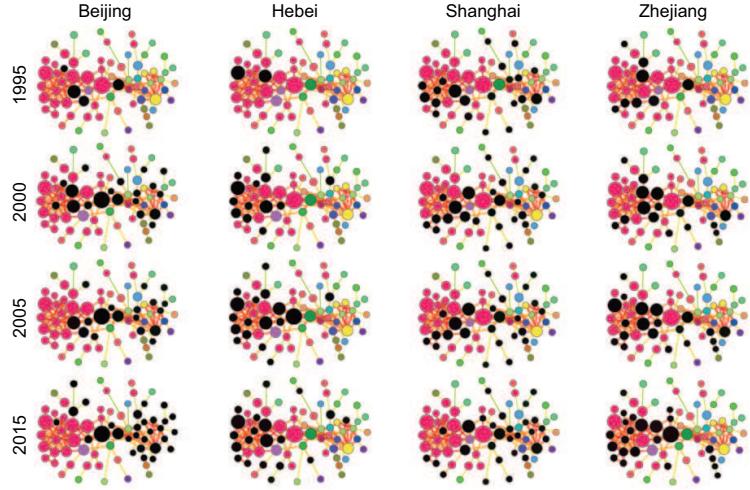


Figure 18: Evolution of China’s regional industrial structure from 1995 to 2015. The industry spaces of four provinces are illustrated, including Beijing, Hebei, Shanghai and Zhejiang. Black circles highlight industries that are present in the industry space of the corresponding province. Figure from [281].

Data from LBSNs have been used to reveal regional economic structure. Based on location information of Weibo users, Liu et al. [266] proposed an effective method to uncover the city-level macro economic structure in China. They employed the linear least square method to model the relations between user number (UN) and GDP for 282 prefecture-level cities. They found that cities below the fitting line are likely to be driven by the tertiary industries, while cities above the fitted line tend to focus on the secondary industry. They quantified the deviation of cities from the fitted line by calculating the measure  $Doo = l_i - y_i$ , where  $l_i$  is the value of fitted line for city  $i$ , and  $y_i$  is the corresponding GDP in the logarithmic form. Further, they used the  $Doo$  measure to predict the macro-economic structure, in particular GDP, by employing the support vector regression (SVR) [279, 280]. They found that the user number (UN) performs better in predicting GDP than some macroeconomic indices such as population and average GDP. This work shows the capacity of online social activity in revealing industrial structure and estimating economic status at the regional level.

Based on the China’s publicly listed firm data from 1990 to 2015, Gao et al. [281] quantified the regional industrial structure of China by constructing a network of related industries, named industry space. First, they estimated the proximity  $\phi_{\alpha,\beta}$  between industries  $\alpha$  and  $\beta$  by calculating the cosine similarity. Formally, let  $x_{i,\alpha,t}$  and  $x_{i,\beta,t}$  be the number of firms in province  $i$  operating respectively in industries  $\alpha$  and  $\beta$  at year  $t$ , the proximity  $\phi_{\alpha,\beta,t}$  is given by:

$$\phi_{\alpha,\beta,t} = \frac{\sum_i x_{i,\alpha,t} x_{i,\beta,t}}{\sqrt{\sum_i (x_{i,\alpha,t})^2} \sqrt{\sum_i (x_{i,\beta,t})^2}}. \quad (35)$$

Then, based on the proximity  $\phi$ , they built the industry space that highlights the relatedness between 70 industries at the sub-sectoral level. The China’s industry space exhibits both a core-periphery structure and a dumbbell structure with a big tightly knit core of manufacturing industries and some small tightly knit cores of service- and information-related activities. Figure 18 presents the evolution of industrial structure of four provinces in China (Beijing, Hebei, Shanghai and Zhejiang) from 1995 to 2015, with black circles showing the industries of presence in the industry space. Specifically, the presence of industry  $\alpha$  in province  $i$  at year  $t$  is identified by the revealed comparative advantage being over 1 (i.e.,  $RCA_{i,\alpha,t} \geq 1$ ) [78]. Beijing and Shanghai gradually occupied Internet and financial services, while Hebei and Zhejiang gradually occupied manufacturing industries. By analyzing the same data, Gao and Zhou [282] found that provinces located along the coast tend to be industrial sophisticated with a high level of economic complexity. Moreover, the provinces’ ranks by their economic complexity are relatively stable during the considered period.

National bureau of industrial enterprises can also be used to portray the production space as the representation of industrial structure. Based on data of four-digit manufacturing sectors from the China’s State Statistical Bureau covering the period of 1999-2007, Guo and He [283] calculated the inter-sector relatedness and produced the pro-

duction space consisting of 424 manufacturing sectors. The production space in 1999 has a core-periphery structure with a major core of electric apparatus, electronic and telecommunications equipment, and a small sub-core cluster consisted of food products, chemical and non-metallic mineral products. The small sub-core cluster developed into an important and dense core of the production space in 2007. They further found that China's regions undergo substantial structural change from 1999 to 2007 with different magnitudes, and industrial evolution has a strong tendency of path dependencies, where regional development is rooted in and subject to the preexisting economic structure.

Economic relatedness contributes significantly to regional industrial diversification. By analyzing Italian trade data during 1995-2003, Boschma et al. [284] presented strong evidence in support of the fact that related variety contributes to regional economic growth. They grouped products into related variety sets  $S_r$  based on the industrial classification and then calculated the related variety index  $RV_i$  of product  $i$  by

$$RV_i = \sum_{r=1}^R P_r H_r, \quad (36)$$

where  $P_r = \sum_i p_i$  is the total exports of region  $r$ , and  $H_r$  is the entropy within the related variety set  $S_r$ . Formally, the entropy  $H_r$  of region  $r$  is given by

$$H_r = - \sum_{i \in S_r} \frac{p_i}{P_r} \log_2 \frac{p_i}{P_r}. \quad (37)$$

They found that a region benefits from extra-regional knowledge originated from related sectors that are already present in that region. Later, Boschma and Frenken [285] demonstrated that technological relatedness affects the process of knowledge spillovers, which benefits regions with different but technologically related activities. As a result, new industries are likely to emerge from related industries. However, the process occurs primarily at the regional level as knowledge spillovers are geographically bounded. Using trade data of 50 Spanish provinces during 1995-2007, Boschma et al. [286] further investigated whether related variety affects regional growth. They calculated two measures of relatedness between industries: the related variety index [284] and the proximity index [77]. They found that Spanish provinces with a variety of related industries exhibit higher rates of economic growth.

By analyzing the US patent data during 1977-1999, Castaldi et al. [287] showed that related technologies can enhance the innovation of a new technology and unrelated variety can enhances technological breakthroughs. Boschma et al. [288] investigated technological relatedness at the city level and technological change in 366 US cities by analyzing the US patent data during 1981-2010. They found that the level of relatedness with existing technologies increases the entry probability of a new technology in a city. Balland et al. [289] discussed the co-evolutionary dynamics between proximity and knowledge ties. They found that proximities might gradually increase due to the past knowledge ties. In particular, the co-evolutionary dynamics can be captured by the processes of learning (cognitive proximity), decoupling (social proximity), agglomeration (geographical proximity), integration (organizational proximity) and institutionalization (institutional proximity). Acemoglu et al. [290] measured the strength of technological flows between technology subcategories using data of 1.8 million US patents and their citation properties during 1975-2004. They found that related pre-existing technological developments have a strong predictive power for future innovations.

A body of literature have demonstrated that relatedness plays an important role in economic development. Indeed, recent empirical evidences have generalized the principle of relatedness [291], which describes the probability that an economy develops or loses an economic activity as a function of the density of its related activities in that economy. Jun et al. [292] studied the role of relatedness in the evolution of bilateral trade. They found that produce relatedness, importer relatedness and exporter relatedness can increase a country's exports of a product. Boschma [293] provided a valuable future research agenda regarding the relatedness as a driver of regional diversification. They suggested to focus on the role of economic and institutional agents. Davids and Frenken [294] recently showed that the type of knowledge being mobilized and produced determines the relative importance of proximity dimensions. They proposed a framework that combines the proximity dimensions with different types of knowledge in the innovation process.

### 3.2.2. Collective learning in economic development

In addition to related varieties, geographic knowledge also plays a crucial role in regional economic development. Boschma et al. [295] demonstrated that capabilities that enable the development of new industries are regional specialized, supporting the hypothesis that knowledge decays strongly with distance in its diffusion process [296]. Due to

the localized nature of knowledge diffusion, neighboring regions should share more similar knowledge and exhibit a geographically correlated pattern in producing structure and economic growth [297]. Scientists have revealed the role of geographic neighbors and highlighted it as an alternative channel for development. Indeed, recent literature have focused on the effects of collective learning—the learning that takes place at the scale of teams, organizations, regions, and nations—by highlighting two learning channels, namely, the inter-industry learning (from related industries), and the inter-regional learning (from neighboring regions)[298, 281].

The effects of geographic knowledge spillovers on firm survival and industry development have been studied based on multiple data at regional, firm and plant levels. Acs et al. [299] analyzed annual data of 11 million establishments in the US private sectors during 1989-1998. By incorporating knowledge spillovers through a geographical variation model, they investigated the relations between regional human capital stocks and new-firm survival. They found that knowledge spillovers lead to higher rates of new-firm survival. Holmes [300] studied the geographic expansion of Wal-Mart stores in the US by analyzing store-level data on sales. They found that locations of new Wal-Mart stores tend to be close to regions where Wal-Mart already had a high density of stores. Broekel and Boschma [301] analyzed data from 59 organizations in the Dutch aviation industry. They uncovered that geographical proximity serves as a driver of network formation and it is a stimulus for firm innovative performance after controlling for the effects of other proximities.

After analyzing a dataset summarizing individual work history, Jara-Figueroa et al. [302] found that the growth and survival of new firms in a location increase when they hire workers with location-specific and industry-specific knowledge instead of occupation-specific knowledge. Moreover, industry-specific knowledge plays a more important role for pioneer than for non-pioneer firms. Using network clustering techniques, Alabdulkareem et al. [303] analyzed the dataset detailing the importance of 161 workplace skills for 672 occupations in the US. They found that skills exhibit a polarization into two clusters: the social-cognitive skills of high-wage occupations and the sensory-physical skills of low-wage occupations. Moreover, workers in occupations relying heavily on one skill cluster are likely to move to other occupations within the same skill cluster, says polarized skill network constrains career mobility of workers.

Based on the international trade data during 1962-2000, Bahar et al. [297] studied the effects of neighboring countries on the evolution of a country's exporting basket. They measured the similarity in countries' export structure by defining an export similarity index (ESI) through the Pearson correlation coefficient. Formally, the ESI between countries  $c$  and  $c'$  is given by

$$S_{c,c'} = \frac{\sum_p (r_{c,p} - \bar{r}_c) \sum_p (r_{c',p} - \bar{r}'_c)}{\sqrt{\sum_p (r_{c,p} - \bar{r}_c)^2 \sum_p (r_{c',p} - \bar{r}'_c)^2}}, \quad (38)$$

where  $r_{c,p} = \ln(\text{RCA}_{c,p} + \varepsilon)$ , and  $\bar{r}_c$  is the average value over all products for country  $c$ . Here,  $\text{RCA}_{c,p}$  is the revealed comparative advantage (RCA) [78] of country  $c$  and product  $p$ , which is calculated by Eq. (11). They found that neighboring countries have significantly larger ESI value than non-neighbors, and ECI is negatively correlated with geographical distance. After using regressions to discount the effects of product relatedness, they further found that a country's probability to export a new product increases significantly (on average, 65% larger) if it has neighboring countries that are already successful exporters of that product.

Based on the survey data of 295 firms in 8 European regions, Broekel and Boschma [304] studied the geographical and cognitive structure of knowledge links. They found that firms' knowledge exchange have differences in their cognitive and geographical dimensions. In particular, connecting with technologically related and similar organizations as well as organizations at various geographical levels (regional and non-regional) can enhance the innovations of firms. By analyzing the data of US state-level exports during 2000-2012, Boschma et al. [305] found that a state in the US has a higher probability (about 58%) of developing a new industry if it has a neighbouring state specialized in that industry. Further, they tested if neighboring regions have more similar export patterns by including ESI given by Eq. (38) into the regression model. They found that the ESI between a pair of states raises by 0.43 standard deviations if the two states share a border in the US.

In a word, previous literature have demonstrated two collective learning channels in regional economic development: the inter-industry learning that involves learning from related industries and the inter-regional learning that involves learning from neighboring regions. Using publicly listed firm data describing the evolution of China's economy between 1990 and 2015, Gao et al. [281] formalized these two collective learning effects. For inter-industry learning, they calculated the density of active related industries ( $\omega$ ) by counting the number of related industries that

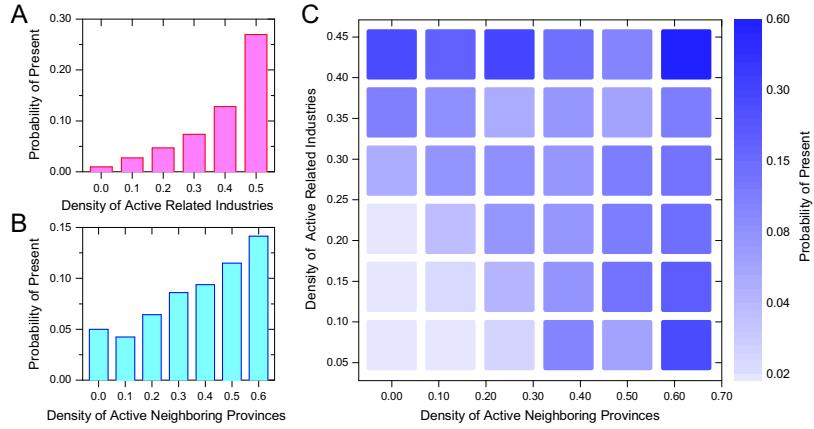


Figure 19: Quantifying the collective learning effects. (A) and (B) are the corresponding marginal probability distributions of new industries present in the next five years, given the density ( $\omega$ ) of active related industries and the density ( $\Omega$ ) of active neighboring provinces, respectively. (C) Joint probability of a new industry developing revealed comparative advantage in a province in the next five years, given  $\omega$  in horizontal-axis and  $\Omega$  in vertical-axis. The color marks the joint probability of new industries present after dividing the two densities into bins. Figure after [281].

are already present (i.e.,  $RCA \geq 1$ ) in that province (see also Ref. [295]). The density  $\omega_{i,\alpha,t}$  for industry  $\alpha$  in province  $i$  at year  $t$  is given by

$$\omega_{i,\alpha,t} = \frac{\sum_{\beta} \phi_{\alpha,\beta,t} U_{i,\beta,t}}{\sum_{\beta} \phi_{\alpha,\beta,t}}, \quad (39)$$

where the binary variable  $U_{i,\beta,t} = 1$  if province  $i$  has advantage in industry  $\alpha$  at year  $t$  (i.e.,  $RCA_{i,\beta,t} \geq 1$ ), and  $U_{i,\beta,t} = 0$  otherwise. They found that the probability for a province to develop a new industry in the next five years increases with  $\omega$  (see Figure 19A), supporting the inter-industry learning effect. For inter-regional learning, they calculated the density of active neighboring provinces ( $\Omega$ ) by counting the number of neighboring provinces that have developed advantage in an industry. The density  $\Omega_{i,\alpha,t}$  for province  $i$  in industry  $\alpha$  at year  $t$  is given by

$$\Omega_{i,\alpha,t} = \sum_j \frac{U_{j,\alpha,t}}{D_{i,j}} \left| \sum_j \frac{1}{D_{i,j}} \right|, \quad (40)$$

where  $D_{i,j}$  is the geographic distance between two provinces  $i$  and  $j$ . They found that the probability that a province will develop a new industry in the next five years increases with  $\Omega$  (see Figure 19B), supporting the inter-regional learning effect.

Furthermore, Gao et al. [281] explored the interaction between inter-regional and inter-industry learning effects. They calculated the joint probability that a new industry will emerge in a province as a function of both densities  $\omega$  and  $\Omega$ . They found that the probability for a province to develop a new industry in a five-year period increases with both  $\omega$  and  $\Omega$  (see Figure 19C). After using a probit model to check the robustness of the results, they demonstrated that the inter-regional and inter-industry learning effects are jointly significant. Interestingly, the regression coefficient of the two densities' interaction term is negative and significant, suggesting the presence of diminishing returns (see Ref. [281] for details). The observation means that, when one learning channel is sufficiently active (inter-industry or inter-regional), the marginal contribution of the other one is reduced. In other words, the two collective learning channels are substitutes for economic development. These empirical findings have been tested and generalized for countries at various stages of development based on different types of data. For example, Gao et al. [306] analyzed over 300 million Brazilian labor records and found evidences in support of the collective learning effects in Brazilian regional economic development.

As geographic knowledge diffusion requires direct forms of human interaction [307], the construction of high-speed rails (HSRs) is likely to facilitate market integration and knowledge spillovers. Using data of China's HSRs, Zheng and Kahn [308] demonstrated that bullet trains help improve the life quality of urban population as HSRs entry allows individuals to access the megacity without living within its boundaries. To explore the impact of HSRs on regional economic activities, Li et al. [309] developed the geographically network weighted regression that incorporates

the changes in network-based travel time from HSRs. They found that HSRs have significantly changed the spatial redistribution of economic activities in regions of China. Later, based on data of prefectural-level cities in China during 1990-2013, Ke et al. [310] explored how HSRs affect the economic growth of cities. They found that the local economic gains are greater for cities connected by HSRs. Meanwhile, Qin [311] found a mild impact of HSRs upgrades on economic growth in China's prefecture-level cities, while the peripheral regions along the upgraded HSRs (e.g., counties close to high-speed rail stations) experienced an investment-driven reduction (3-5%) in GDP and GDP per capita after 2007.

The effects of modern transportation (e.g., HSRs and flights) on economic development and knowledge spillovers have also been studied in developed countries. For the European Union, Kim et al. [312] explored the contribution of HSRs in promoting economic integration. They found that local economic development is necessarily led by transport improvements alone, especially when this involves cross-border links. By analyzing the northwest European HSRs and the UK's first HSR, Vickerman [313] found that transport infrastructure by itself does not likely have a transformative effect on economy, but it can contribute to such effect after being coupled with policy interventions such as policies related to complementary planning and policies towards labour markets. Ahlfeldt and Feddersen [314] analyzed the economic impact of the German HSR and found that HSR has a causal effect (on average about 8.5%) on GDP growth in the regions of intermediate stops. Moreover, the strength of spillovers halves every 30 minutes of travelling time and diminishes to zero after about 200 minutes. Besides HSRs, a reduction in travel cost brought by cheaper flights can also facilitate knowledge spillovers reflected by scientific collaborations. Catalini et al. [315] analyzed a scientist-level dataset covering all US chemistry faculty members during 1991-2013. They found that scientific collaborations increase by 50% after the Southwest Airlines opens a new route, showing that face-to-face interactions can enhance scientific collaborations.

To address endogenous concerns of inter-regional learning, Gao et al. [281] applied the differences-in-differences (DID) analysis [165] and used the introduction of HSRs as an adequate instrument. The underlying intuition is that HSRs entry reduces the barriers to the inter-regional learning but should not affect the inter-industry learning. Specifically, they used the DID analysis to test whether provinces connected by HSRs increased their industrial similarity and experienced a boost in the productivity of shared industries. The industrial similarity  $\varphi_{i,j,t}$  between a pair provinces  $i$  and  $j$  at year  $t$  is measured by

$$\varphi_{i,j,t} = \frac{\sum_{\alpha} y_{i,\alpha,t} y_{j,\alpha,t}}{\sqrt{\sum_{\alpha} (y_{i,\alpha,t})^2} \sqrt{\sum_{\alpha} (y_{j,\alpha,t})^2}}, \quad (41)$$

where  $y_{i,\alpha,t} = \ln(\text{RCA}_{i,\alpha,t} + 1)$  and  $y_{j,\alpha,t} = \ln(\text{RCA}_{j,\alpha,t} + 1)$ . They found that the industrial similarity ( $\varphi_{i,j}$ ) decays strongly with the geographic distance ( $D_{i,j}$ ), and HSRs entry significantly increases the industrial similarity between provinces connected by HSRs. Moreover, the labor productivity (measured by the revenue per worker) increases in the provinces connected by HSRs, supporting the hypothesis that HSRs entry promotes inter-regional learning.

### 3.2.3. Development paths and strategies

Regional industrial diversification has been suggested as a strong path-dependent process, where economic relatedness plays a significant role. Based on plant-level data of 70 Swedish regions during 1969-2002, Neffke et al. [81] identified related industries using the revealed relatedness (RR) measure [316]. They found that the probability that an industry will enter (exit) a region increases (decreases) with the number of related industries already present in that region. Neffke et al. [317] further studied the effects of technological relatedness on plant survival in Sweden during 1970-2004. They found that the plant survival rates are increased by the presence of technologically related local industries. Further, Neffke and Henning [318] investigated how industry's skill relatedness affects the diversification of firms by calculating the RR measure based on the labor flow data covering about 4.5 million workers in 400 industries in Sweden during 2004-2007. They found that firms tend to diversify into industries that require skills strongly related to the firms' existing industries. These works suggest the predictive power of skill relatedness for firm diversification.

Some literature have also explored the role of relatedness in regional development, industrial structural change and firm survival in China [319]. Howell et al. [320] analyzed the data of over 13 million entrepreneurial firms in China during 1998-2007. They found that local related variety has a stronger positive effect than other types of agglomeration on new firm survival. Moreover, the intensity and location of governmental support affect post-entry performance and survival of firms. He et al. [321] analyzed the annual survey of industrial firms in China during 1998-2005. They found that private enterprises rely more on market-oriented institutions, while firms with local

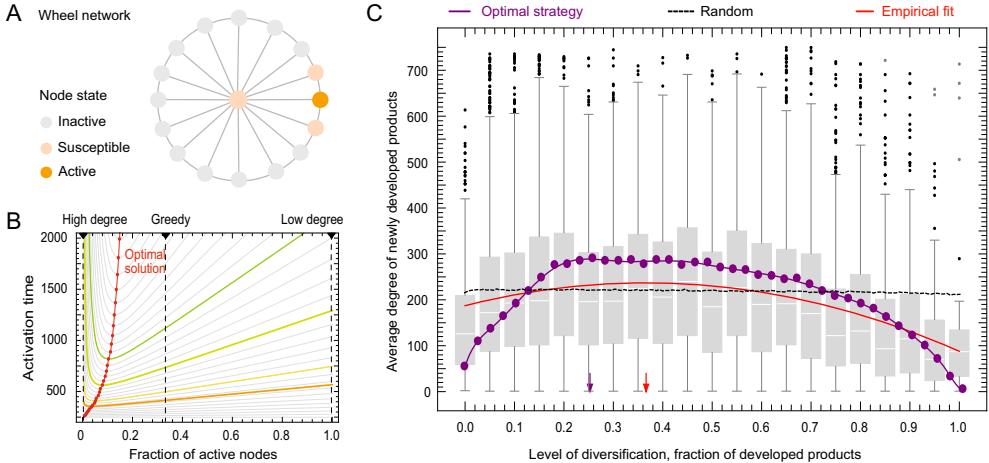


Figure 20: Strategic diffusion in networks, and the diversification of products. (A) A wheel network. (B) Time needed to activate all nodes in the wheel network as a function of the time when the hub is targeted (i.e., the fraction of active nodes). The red line indicates the optimal time based on the strategic diffusion. (C) Box-plot presenting the average degree of newly developed products in the product space as a function of the fraction of developed products. The red line shows the fit for the empirical values, and the purple line shows the optimal mix of greedy and high degree strategies obtained via numerical simulations. The black line shows the null model baseline which uses the random strategy. Figure after [326].

governmental supports and industrial linkages are more likely to sustain. He et al. [322] later analyzed firm-level data of manufacturing industries in China during 1998-2008. They found that regions tend to develop new industries that are technologically related to the existing portfolio. These results demonstrate that regional industrial development is a path-dependent process where industries related to pre-existing ones.

Recently, Gao [323] investigated how to maximize the learning from related industries (i.e., inter-industry learning) and neighboring regions (i.e., inter-regional learning) by leveraging the Brazilian labor data. He used a simple variant of the threshold model [324] to simulate the diversification of industries on real networks. In the threshold model, a region or an industry will be activated if over half of its neighbors are already active [325]. For inter-regional learning, simulations are based on the Brazilian industry space [306], and the set of initial industries are selected according to a turnable balancing index of core and periphery industries. Gao [323] found an optimal strategy that results in a good tradeoff between core and periphery industries in the initial activation. For inter-regional learning, simulations are based on the adjacent network of regions integrated with one spatial link being added between each pair of regions [325], and the set of initial industries are randomly selected. The lengths of spatial links are determined by a turnable balancing index of nearby and distant regions. The result suggests an optimal strategy that makes a balance between nearby and distant regions in establishing new spatial connections. These findings demonstrate that there are optimal strategies for both channels that can maximize the learning effects in industrial diversification.

As suggested by many empirical studies, countries and regions are likely to develop economic activities that have close relatedness to what they have already developed, yielding the principle of relatedness [291]. In other words, the probability of developing a new industry in a region increases with the density of the region's developed industries that are related to the industry. As the produce space [77] and industrial space [281, 323] have a core-periphery structure, the difficulty and opportunity of developing produces and industries at different locations of the space are different. Alshamsi et al. [326] explored the optimal diversification strategies in the produce space (see Figure 20). They showed that the high-degree strategy, i.e., always targeting the potentially products with the highest degree (e.g., products in the core), will result in a long activation time, while the low-degree strategy, i.e., always targeting the potentially products with the lowest degree (e.g., products in the periphery), will miss the opportunity for a rapid development. In order to minimize the total time needed to develop all products, they proposed a method named strategic diffusion to identify products that are optimal to target at each time step. The optimal strategy targets core produces during a narrow and specific time window, which comes earlier than we previously thought (e.g., the time by the greedy strategy). They analyzed the international trade data and demonstrated that the countries' strategies to diversify their products are close to the optimal ones. The time that countries target core products, however, is later than the one

suggested by the model, showing that countries can still save the total time of developing all products.

The path-dependent process of regional diversification suggests that regions have more opportunities to develop industries that have high relatedness to their pre-existing ones [293], while the strategic diffusion suggests that countries can optimize their development paths by targeting highly connected but somewhat unrelated activities at a certain time [326]. The development of unrelated economic activities is particularly significant for the catching-up growth in developing economies as it is usually hard for them to jump from periphery to core areas in the product and industry space, say the space conditions the development [77]. Regarding this point, Zhu et al. [327] explored the development paths of regions in the heterogeneous industry space built on the exporting data of Chinese firms during 2002-2011. They studied whether developing regions can catch up by breaking the path-dependent trajectories and jumping farther into core areas of the uneven industry space. They demonstrated that developing regions can make a farther jump to new industries in a path-breaking way, and the reliance of technological relatedness can be transcended by internal innovations and extra-regional linkages. These findings suggest that less developed economies should pay more attention to improving other factors (such as infrastructure and education, government supports and extra-regional linkages) to promote their jumping capability in the catching-up growth.

Processes of unrelated diversification are also important for economic development, and economies can benefit from entering unrelated activities. Boschma et al. [328] argued that a theory of regional diversification should also accounts for the processes of unrelated diversification. They suggested to pay attention to the role of agency in institutional entrepreneurship and enabling factors at different spatial scales. In particular, they discussed four regional diversification trajectories including two related diversification (replication and exaptation) and two unrelated diversification (transplantation involves and saltation stands). Pinheiro et al. [329] identified the periods that countries entered unrelated products by analyzing the diversification paths of 93 countries in product exports during 1965-2014. They found that countries tend to enter unrelated products when they have high levels of human capital and during their intermediate level of economic development. Moreover, countries that entered more unrelated products experienced a significant increase in economic growth, showing the positive gain to target unrelated activities at a specific development stage. All the above results indeed ask for more intelligent strategies for economic diversification by balancing related and unrelated activities in development.

### 3.3. Urban scalings and perception

The availability of large-scale and quantitative data from socioeconomic systems and image database has enhanced our perception of urban landscape and surrounding environment. In this subsection, we will summarize empirical observations and theoretical explanations of scaling laws of urban population with urban metrics (e.g., crime rate, employment, innovation and economic activity). Then, we will review recent applications of novel data on inferring the function of urban areas. Next, we introduce crowdsourcing methods and computational vision techniques to measure livability, safety and inequality, to infer the status of urban life, and to quantify the changes of urban streetscapes. Finally, we will introduce recent progresses on urban computing for better development in urban areas.

#### 3.3.1. Scaling laws for cities

Empirical observations in economics suggest the Zipf's law for cities in most countries. That is, the number of cities with populations greater than  $N$  is proportional to  $1/N$ . Formally,  $P(\text{size} > N) = Y_0/N^\beta$ , with  $\beta \approx 1$  and  $Y_0$  being a constant. Gabaix [330] provided a simple explanation for the emergence of such Zipf's law. They demonstrated that the power-law exponent  $\beta = 1$  is necessarily led by the most natural conditions on the Markov chain. Using the maximum likelihood estimation (MLE) method, Clauset et al. [331] estimated the power-law exponent for the populations of US cities in 2000, finding a rank-size slope of  $-0.73$  (i.e.,  $\beta = 0.73$ ). Later, Small et al. [332] tested the Zipf's law based on a unique proxy for anthropogenic development, specifically, the temporally stable nighttime lights (NTLs) from the DMSP-OLS. They found that the estimated  $\beta$  ranges from 0.95 to 1.11, suggesting that Zipf's law holds for spatial extent of anthropogenic development at global scales.

Urban scaling laws provide a quantitative connection between urbanization and economic development, which is common to all cities around the world. By analyzing datasets from urban systems in the US, Germany and China, Bettencourt et al. [333] found that many diverse urban variables fit power-law functions of population size with scaling exponents  $\beta$  falling into distinct universality classes. Using total population  $N(t)$  to estimate the size of the city at time  $t$ , the power-law scaling takes the form

$$Y(t) = Y_0 N(t)^\beta, \quad (42)$$

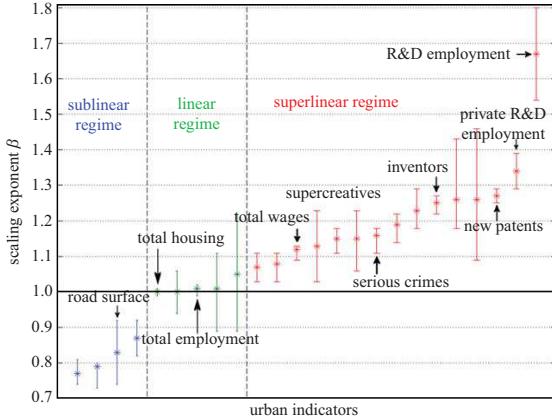


Figure 21: Scaling exponents for urban metrics versus city size. Scaling exponents  $\beta$  found for China, Germany and USA are shown with 95% confidence interval for different urban indicators. Scaling exponents are colour-coded according to their regime: sublinear in blue, linear in green, and superlinear in red. Figure from [334] with data coming from [333].

where  $Y(t)$  denotes a certain metric on social activities or material resources at time  $t$ , and  $Y_0$  is the normalization constant. They found a pervasive property of urban organization with exponents falling into three categories (see Figure 21 for the results summarized by Arcaute et al. [334]):  $\beta > 1$  (superlinear),  $\beta = 1$  (linear), and  $\beta < 1$  (sublinear). In particular, they showed that  $\beta \approx 1$  is usually associated with individual human needs such as housing, employment and household electrical consumption,  $\beta \approx 1.2 > 1$  is associated with social currencies such as information, innovation and wealth, and  $\beta \approx 0.8 < 1$  is associated with infrastructure such as road surface, gasoline stations and length of electrical cables.

Urban scaling laws have been widely observed in emissions and infrastructures. Louf and Barthélémy [335] found that the CO<sub>2</sub> emission scales superlinearly with  $N$  in the US in 2012 ( $\beta = 1.26$ ) and the OECD countries in 2008 ( $\beta = 1.21$ ). Oliveira et al. [336] analyzed data of CO<sub>2</sub> emissions in the US during 1999–2008 and found a superlinear scaling (with an average exponent  $\langle \beta \rangle = 1.46$ ) across all cities. Delong and Burger [337] found that energy use scales superlinearly with  $N$  in Sweden, England and Wales (E&W), the US, and the world. Samaniego and Moses [338] analyzed the structure of road networks in 425 US cities. They found that road network capacity per capita is independent of city size measured both by population and spatial extent of the urban area. Batty [339] analyzed the road network of cities in E&W and found a superlinear scaling ( $\beta \approx 1.09$ ) of road accessibility with  $N$ . Louf et al. [340] analyzed the data of about 140 subways and over 50 railway networks across the world. They found that the length of subway networks scales superlinearly ( $\beta \approx 1.13$ ) while the yearly ridership of railway networks scales linearly with the number of stations. For the UK and urban California, Masucci et al. [341] found that both the total length  $L(N)$  and the area  $A(N)$  of street networks scale almost linearly with  $N$ , and the urban scalings persist in space and time.

A body of literature have demonstrated the urban scaling of crime in cities. Alves et al. [342] analyzed data of homicides in Brazilian cities and found that the number of homicides scales superlinearly ( $\beta \approx 1.15$ ) with  $N$ . They further proposed an approach to unveil relations between crime and urban metrics using the distance between the actual homicide number and the expected number from the scaling law. Banerjee et al. [343] analyzed the data of US cities and found that crime scales superlinearly ( $\beta = 1.26$ ) with  $N$ . They gave the explanation that the number of police scale sublinearly while the number of generated crimes scales linearly. After analyzing monthly police crime reports in E&W, Hanley et al. [344] found four types of scaling behaviors based on population density: non-urban scaling, accelerated scaling ( $\beta_L < \beta_H$ ), inhibited scaling ( $\beta_L > \beta_H > 0$ ) and collapsed scaling ( $\beta_L > \beta_H$ , with  $\beta_H < 0$ ), where  $\beta_L$  and  $\beta_H$  are the scaling exponents for low and high population density, respectively. Oliveira et al. [345] analyzed the disaggregated criminal data from the US and UK. They found that the crime concentration does not scale with the city size, and the crime distribution in a city follows a power-law distribution with exponent depending on the crime type.

In most of these aforementioned literature, the word “city” refers to a larger agglomeration around the central

city, which is socioeconomic unit instead of administrative definition. In fact, there are alternative definitions of city boundaries. Arcaute et al. [334] developed a framework to produce a system of cities by clustering small statistical units. They found that the scaling exponent  $\beta$  gives mild deviations from linearity in E&W, suggesting that economic intricacies are not fully grasped by the urban population  $N$ . Van Raan et al. [346] analyzed the urban scalings in the Netherlands and found a superlinear scaling ( $\beta \approx 1.15$ ) of GDP with  $N$  for major cities. After considering three separate modalities, they found that municipalities perform better than urban agglomerations and urban areas with the same population, showing that cities with a municipal reorganization are likely to perform better. Bettencourt and Jose [347] applied new harmonized definitions of functional urban areas to examine scalings, finding that pooling together cities from different urban systems can better identify scaling behaviors in European cities.

Social ties of cities also exhibit scaling behaviors. Pan et al. [348] found that the density of social ties  $T(\rho)$  scales superlinearly with urban population density  $\rho$ . The social-tie density is given by  $T(\rho) = \rho \ln \rho + (C - 1)\rho$ , where  $C = 2 \ln r_{\max} + \ln \pi + 1$  with a unique  $r_{\max}$  for each city. In particular,  $\beta$  falls into a narrow band  $1.1 \leq \beta \leq 1.3$ , where  $\beta = 1.21$  for the AIDS/HIV prevalence in the US cities and  $\beta = 1.26$  for the total GDP per square kilometer in the European cities. Moreover, the superlinear scaling ( $\beta > 1$ ) is led by the increase in  $\rho$ , and the diffusion rate along social ties can accurately reproduce urban scalings. After analyzing mobile phone data of 31 Spanish cities, Louail et al. [349] found that the number of activity centers scales sublinearly ( $\beta < 1$ ) with  $N$ . Markus et al. [350] analyzed the nationwide communication records in Portugal and the UK. They found that the total number of contacts and communication activities scale superlinearly with  $N$ . Recently, Leitão et al. [351] studied the existence of nonlinear scaling by developing a statistical framework to account fluctuations. They found that  $\beta$  does not only depend on the fluctuations contained in the datasets but also on the assumptions of models and the heavy-tailed distribution of city sizes.

Several explanations have been proposed for the origin of urban scalings. Arbesman et al. [352] explained the observed superlinear scaling in the relations between population size and innovation by a network model, where the number of long-distance ties associated with a city is proportional to its population and these ties provide the potential for innovation. The model yields a reasonable range of the scaling exponent, suggesting socially distant ties as a powerful force of the superlinear scaling. Later, Gomez-Lievano et al. [353] built a statistical framework to explore how urban scaling laws emerge and relate to Zipf's law. Using data of homicides in three cities, they derived the conditional probability density  $P(Y|N)$  for the number of homicides  $Y$  in a city with population  $N$  by exploiting the Bayes' rule

$$P(Y|N) = \frac{P(N|Y)P(Y)}{P(N)}, \quad (43)$$

where  $P(Y)$  is the distribution of homicides in cities, and  $P(N|Y)$  is the conditional probability for the populations of cities with a given number of homicides. After studying the statistical properties of  $P(Y)$  and  $P(N|Y)$ , they found that scaling laws emerge as the expectation value of  $Y$ , which is a function of  $N$ . Moreover, the knowledge of the distribution  $P(Y)$  can be used to predict the Zipf's exponent from the statistics of urban metrics.

To better understand the origin of urban scalings, Bettencourt [354] developed a framework to estimate scaling exponents without modeling infrastructure. In a city with land area  $A$  and population  $N$ , the strength of local interactions between people in an area  $a_0$  is denoted as  $g$ . The basic ideas behind their model are summarized as follows. First, the number of local interactions per person is given by  $a_0 \ell \cdot N/A$ , where  $N/A$  is the population density, and  $\ell$  is the length of travel. Then, a city's total social output  $Y$  is given by

$$Y = \bar{g} \cdot N \cdot a_0 \ell \cdot \frac{N}{A} = G \cdot \frac{N^2}{A}, \quad (44)$$

where  $G \equiv \bar{g} a_0 \ell$ ,  $\bar{g}$  is the average social output per interaction, and  $N$  is the population size. Next, the total cost to mix the city is  $T = \varepsilon L N = \varepsilon A^{1/2} N$ , where  $\varepsilon$  is a force per unit time, and  $L = A^{1/2}$  is the cost per person. The cost should be covered by each individual,  $y = Y/N$ , requiring  $y \approx T/N$ . This implies  $A(N) = aN^\alpha$  with  $\alpha = 2/3 < 1$  (sublinear scaling) and  $a = (G/\varepsilon)^\alpha$ . Thus, they obtained  $Y = Y_0 N^\beta$ , where  $\beta = 2 - \alpha = 4/3 > 1$  (superlinear scaling) and  $Y_0 = G^{1-\alpha} \varepsilon^\alpha$ . Yakubo et al. [355] provided explanations for both scalings based on a geographical scale-free network. The individual activity  $y_{ij}$  of two connected nodes  $i$  and  $j$  depends on their Euclidean distance  $l_{ij}$ . The urban metric  $Y(N)$  scales superlinearly when  $y_{ij}$  increases with  $l_{ij}$  (e.g., for creative productivities),  $Y(N)$  scales sublinearly when  $y_{ij}$  decreases with  $l_{ij}$  (e.g., for infrastructures), and  $Y(N)$  scales linearly when the geographical constraint is strong enough.

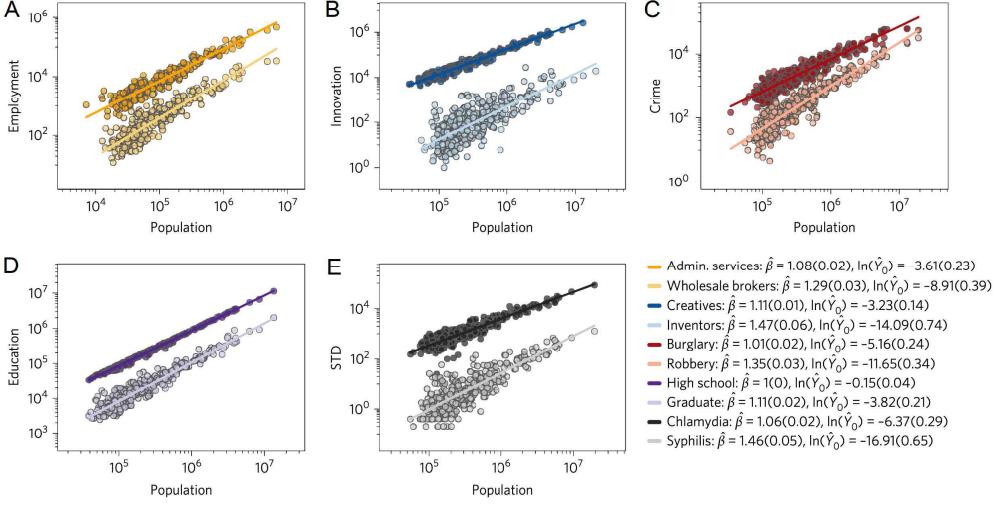


Figure 22: Explanations of ten different urban scaling phenomena. (A) Employment in two industries. (B) Two types of innovative activities. (C) Two types of violent crime. (D) People with a given educational level. (E) Two sexually transmitted diseases (STDs). Lines represent the best fit of the model  $E\{Y|N\} = Y_0 N^\beta$ . Hat ( $\hat{\cdot}$ ) denotes a statistical estimate of a parameter. Figure from [356].

Recently, alternative explanations of urban scalings have been suggested. Gomez-Lievano et al. [356] showed that the number of people engaged in each phenomena scales as a power law with population size (see Figure 22). Accordingly, they proposed a simple model to explain urban scalings, where each phenomenon depends on a number of factors  $M$ , an individual requires one factor with probability  $q$ , and a city provides one factor with probability  $r$ . The aggregated output of a given phenomenon is modeled by  $Y = \sum_{j=1}^N X_j$ , where  $N$  is the population size,  $X_j = 1$  if individual  $j$  gets all required  $m$  factors, and  $X_j = 0$  otherwise. If  $m$  factors have been required, an individual will not require the other  $M - m$  factors. Given a city with  $m$  factors, the probability of an individual involved in the activity is  $P\{X_j = 1|M_{\text{city}} = m\} = (1 - q)^{M-m}$ , where  $M_{\text{city}} = B(M, r)$  is a binomially distributed random variable. The expected value of  $Y$  is given by

$$E\{Y\} = N \sum_{m=0}^M P\{X_j = 1|M_{\text{city}} = m\} P\{M_{\text{city}} = m\} \approx NP, \quad (45)$$

and the variance of  $Y$  is given by

$$\text{Var}\{Y\} \approx E\{Y\}^2 \left( \frac{1}{E\{Y\}} - \frac{1}{N} + \frac{1}{p^q} - 1 \right), \quad (46)$$

where  $P \equiv e^{-Mq(1-r)}$  (see Ref. [356] for details). By assuming that  $r = a + b \ln(N)$ , the scaling function  $E\{Y\} = Y_0 N^\beta$  can be obtained. The model suggests that phenomena requiring more factors will scale more superlinearly (i.e., with larger  $\beta > 1$ ). Ribeiro et al. [357] proposed an explanation of urban scalings based on the interactions between individuals and the fractal dimension of cities. Their framework can reproduce the urban scaling for infrastructure (sublinear) and social indicators (superlinear). Using a spatial attraction and matching growth mechanism, Li et al. [358] proposed a unified model that can reproduce the spatial scalings for population, total road length, and total number of socioeconomic interactions. Their model presents consistent results with empirical data and explains the origins of sublinear and superlinear scalings.

Urban scaling laws have been applied to differentiate urban economic productivity and quantify intrinsic diversity of urban economic activities. Based on the deviations from general scaling laws, Bettencourt et al. [359] proposed new metrics of a city's dynamics and urban performance (e.g., personal income, patents and violent crime). Formally, the deviation  $\xi_i$  of a metric is quantified by the residuals [360]

$$\xi_i = \log \frac{Y_i}{Y(N_i)} = \log \frac{Y_i}{Y_0 N_i^\beta}, \quad (47)$$

where  $Y_i$  is the observed value of metric  $i$  for an arbitrary city,  $Y(N_i)$  is the average value of urban metrics, and  $N_i$  is the population size. The scale-adjusted metropolitan indicator (SAMI)  $\xi$  is dimensionless and independent of  $N$ . Moreover, SAMI captures the specific dynamics of a city and represents its performance relative to other cities. This method provides a promising way to rank cities without the population size bias. Lobo et al. [361] derived a new expression for the total factor productivity (TFP) of urban areas. The scale-adjusted urban TFP is well-approximated by  $\xi_i^A \approx (1 - \alpha)(\xi_i^W - \xi_i)$ , where  $\alpha$  is the production factors,  $\xi_i^W$  is the SAMI for total labor income, and  $\xi_i$  is the SAMI for total capital income. They found a systematic dependence of urban productivity on population size. Youn et al. [362] analyzed records of establishments in the US urban areas and found that the total establishment number scales linearly with the city size. Further, they proposed a framework to measure the intrinsic diversity of economic activities, revealed the universal scaling distribution of business types, and presented a simple mathematical derivation of the universality.

### 3.3.2. Unfolding urban functional areas

In regional and urban economic development, a variety of data from remotely sensing (RS), mobile phones (MPs) and online social media (SM) have been used to map the function of regions and capture the urban structure. In the following, we will introduce the applications of nighttime lights (NTLs) data from the DMSP-OLS to measure the spatio-temporal urban dynamics. Then, we will review literature that leveraged novel data sources (e.g., MPs, Twitter, and check-ins) to uncover the inherent characteristics of functional regions and to predict regional economic status. At last, we will summarize recent progresses on the analysis and prediction of house price based on the data of satellite imagery and subway in addition to the traditional market data.

NTLs data have been used to monitor the dynamics of urban structure. Sutton [363] developed a measure of urban sprawl based on NTLs imagery. The urban sprawl is scale-adjusted to an urban area's total population, and the areal extent of metropolitan areas is measured based on the NTLs of the US. They found that inland and midwestern cities have more urban sprawl than west coast cities. This work sheds some insights to the spatial patterns of urban sprawl that is difficult to be precisely defined. Pandey et al. [364] extracted urban areas and monitored urbanization dynamics of India based on NTLs data from the DMSP-OLS and the SPOT vegetation [365]. They employed the SVM-based classification algorithm to extract urban land extent from NTLs (see also Ref. [366]) and verified the results by global urban extent map and Google Earth images. The state-wise increase in urban area is consistent with the change in urban population and GDP, showing the applicability of NTLs to quantify urban patterns.

Based on NTLs data, Zhang and Seto [367] monitored urban changes at the regional scale. They applied an iterative unsupervised classification method to analyze the NTLs data from the DMSP-OLS during 1992-2008 and mapped urbanization dynamics in China, Japan, India and the US. They found that India had higher growth rates than China between 1992 and 2000, while China experienced higher rates of urban growth than India between 2000 and 2008. Frolking et al. [368] analyzed the data from the SeaWinds microwave backscatter power return (PR) [369] and the DMSP-OLS. They found different evolution patterns of urban structure between India and China. Chinese cities rapidly expanded their built-up infrastructure in both height and extent with increased urban population density. By comparison, Indian cities were primarily built out with larger areas of lower population density. These results suggest two distinct trajectories of urban growth.

Recently, Li et al. [370] analyzed the spatio-temporal urban dynamics by applying a linear regression method to time-series from the DMSP-OLS for the southeast US. They found that the newly built urban areas can be effectively detected, while the urban expansion cannot be explained solely by population growth. Huang et al. [371] mapped sub-pixel urban expansion of Chinese cities based on the data from the DMSP-OLS and the Moderate Resolution Imaging Spectroradiometer (MODIS) [372]. They applied the random forest regression model to estimate sub-pixel urban percentage with the high quality calibration information derived from the Landsat data. The estimation of urban land area can be improved by including data from the MODIS and DMSP-OLS. Based on a set of landscape metrics, Liu et al. [373] explored the general spatio-temporal patterns of urbanization by examining 16 world-wide cities during 1800-2000. They uncovered several common urbanization patterns. For example, urban landscape becomes more fragmented, diverse and complex in the urbanization process.

MP data have been used to estimate the characteristics of functional regions. Based on a MP dataset consisting of 431 million calls and the involved locations of mobile base towers, Chi et al. [374] constructed a cell-based spatially embedded interaction network of regions in Heilongjiang province, China. The cells (Voronoi polygons) are used to approximate the service area of a mobile base tower. They calculated the betweenness centrality [375] of a cell  $k$  in

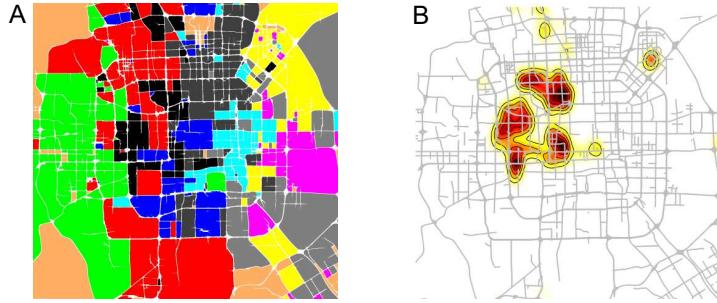


Figure 23: Identification of functional regions in a city. (A) Functional regions. Regions have similar functions are filled with the same color. (B) Intensity of a function. The functionality intensity of developed commercial/entertainment (a kind of function) areas in Beijing is illustrated. The darker part suggests a higher intensity. Figure from [378].

the unweighted interaction network by

$$C_B(k) = \sum_{i \neq j \neq k}^N \frac{\sigma_{ij}(k)}{\sigma_{ij}}, \quad (48)$$

where  $\sigma_{ij}$  is the number of shortest paths from nodes  $i$  to  $j$ , and  $\sigma_{ij}(k)$  is the number of shortest paths passing through node  $k$ . They found that cells with high  $C_B$  are distributed linearly across the province. After applying a community detection algorithm [376], they found a two-level hierarchical organization embedded in the interaction network, where the bottom-level communities respect the county boundaries, and the top-level communities respect the prefecture-level unit boundaries. Moreover, almost every community has a cell with high  $C_B$  at the commercial center or the government seat. Toole et al. [377] studied dynamic land usages based on the temporal activity patterns of MP users. They demonstrated that supervised classification of labeled MP zoning data exhibits reasonable accuracy in identifying clusters of locations.

Different functional regions in urban areas can be identified by data about points of interests (POIs) and human mobility. Yuan et al. [378] proposed a framework named DRoF (Discovers Regions of different Functions) to discover regions of different functions in a city (see Figure 23). They segmented a city into disjointed regions by major roads and employed a topic-based model to infer the functions of each region. The topic-based model treats a urban region as a document, regards mobility patterns as words, deems a function as a topic and uses POIs as metadata. In this way, they represented a region by a distribution of functions and denoted a function by a distribution of mobility patterns. Next, they clustered regions by the topic distribution and identified the intensity of each function. Experiments based on POI datasets and taxicab-based GPS trajectory datasets of Beijing demonstrated that their method outperforms baseline methods solely using POIs or human mobility. Further, Yuan et al. [379] extended the DRoF framework by introducing the concept of latent activity trajectory (LAT) to capture citizens' socioeconomic activities. Their new method employs a morphological approach to segment a city and applies a collaborative-filtering-based approach to learn the location semantics from POIs. Their method exhibits improved performance in identifying functional zones using location and mobility mined from semantics LAT.

Data from location-based social networks (LBSNs) have been used to identify land usages and derive commercial locations. Based on Twitter's spatial (geotagged) and temporal (time-stamped) data, Frias-Martinez et al. [380] presented methods to automatically determine land usages and locate urban POIs. Preliminary validation in Manhattan suggests geotagged tweets as a powerful data source for characterizing urban landscapes. Frias-Martinez et al. [381] further used unsupervised learning to cluster regions with similar patterns of tweeting activities. They verified the new method in Manhattan, London and Madrid based on tweeting activities and ground truthing information for land usages. Recently, Lloyd and Cheshire [382] estimated locations of retail centers from geotagged tweets. They used an adaptive kernel density estimation to identify retail-related tweets and examined their spatial attributes. They showed that areas of elevated retail activities can be well located by retail-related tweets. Soliman et al. [383] analyzed 39 million geotagged tweets in Chicago and found that the majority's temporal patterns of tweeting at key locations are significantly associated with the types of land usage. They proposed a novel approach that can classify key locations into types of land usage with an overall accuracy of 0.78.

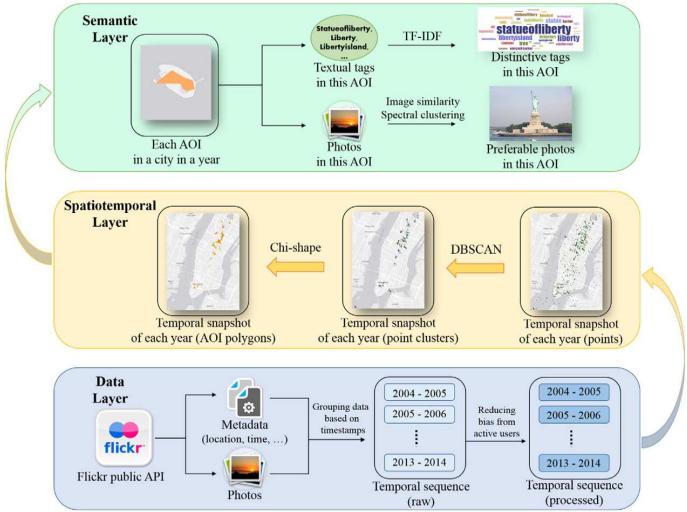


Figure 24: A three-layer coherent framework for extracting urban areas of interest (AOI) from geotagged Flickr data. The framework covers data pre-processing, point clustering, area construction, and semantics enrichment. Figure from [387].

Check-in data have also been used to analyze regional structure and predict urban economic status. Shen and Karimi [384] analyzed check-in data to explicitly portray urban structure. They proposed a novel framework to characterize urban streets with land-use connectivity indices and introduced a model to package three principal dimensions of an urban function network into one integrated index (see Tianjin in China as an example in Ref. [384]), which can explicitly describe the inherent function structure and the regions' typology across scales. Zhi et al. [385] proposed a model to infer functional regions from about 15 million check-in records during a year-long period in Shanghai, China. They used a novel low-rank approximation model to identify a series of latent spatio-temporal activity structures and obtained a series of underlying associations between the spatial and temporal activity patterns. The model is applicable to estimate functional regions including commercial dominant areas, developed residential areas, and developing residential areas.

Urban areas of interest can also be extracted from online volunteered geographic information such as VGI, POIs, and geotagged photos. By utilizing VGI-based POI data obtained from Yahoo! in part of the Boston metropolitan area, Jiang et al. [386] applied a local maximum likelihood estimation to determine disaggregated land usage. They showed that employment estimations based on VGI-based POI data (Yahoo!) match estimations based on proprietary business establishment databases. Their method provides an alternative to estimate disaggregated land usages in a timely manner as POI data can be obtained at a high frequency with a low cost. Based on geotagged photos from Flickr, Hu et al. [387] developed a coherent framework to extract and understand urban areas of interest (AOI). Their framework covers data pre-processing, point clustering, area construction, and semantics enrichment (see Figure 24). They applied the method called density-based spatial clustering for applications with noise (DBSCAN) [388] to identify AOI, employed the method named term frequency and inverse document frequency (TF-IDF) [389] to extract distinctive textual tags for understanding AOI and designed a workflow to select photos capturing a preferable view of an AOI. Their framework provides a better identification of AOI and helps understand dynamics of AOI.

Affected by location and neighborhood in regional and urban areas, house price is a sensitive and important signal in real estate market and even regional economic development. Many methods have been proposed to analyze and predict house price, for example, the hedonic house price model [390]. After analyzing purchase records from housing market in Chengdu, China, Xin and Zheng [391] found that the spatial structure in housing data sets is important for house price estimation. They proposed a spatial hedonic model and successfully applied it to estimate house price index at the urban zone level. Zhang et al. [392] studied how urban village removal affects nearby house price in Beijing, China. They found that an average urban village discounts house price by 2.5%, while its removal increases house price by over 3.3%. Zheng et al. [393] studied the impact of subway transit on local house price by analyzing data of subway construction history, restaurant activities in station neighborhoods and rental housing transactions

in Beijing, China. They found that changes in restaurant activities capture 20-40% of house price appreciation in neighborhood of new subway stations.

Online searches and satellite imagery data have also been used to track urban housing market dynamics. Based on Google search activity data, Zheng et al. [394] constructed a real estate confidence index to measure the view on future housing market trend. They tested how the confidence index is associated with prices of newly built housing units in 35 Chinese cities during the past 10 years. The confidence index can predict the sale of new houses and the increase of local house price, while it has heterogeneous impacts on local real estate outcomes. Recently, Bency et al. [395] proposed a deep learning framework to predict house price based on satellite imagery. They extracted features from satellite imagery by training deep convolutional neural networks (DCNNs) [20] to capture the neighborhood effects and then trained multiple models to regress house price using the extracted features. After validating results based on POI data, they demonstrated that leveraging neighborhood information embedded in satellite images can improve the accuracy of house price prediction.

### 3.3.3. Perceiving urban environment

Visual appearances of urban spaces are thought to have significant effects on psychological states of inhabitants, behaviors of citizens, and socioeconomic outcomes in neighborhoods. Recently, high spatial resolution data have been used to quantify the perception of urban environment and its relation to residents' health. The Google Street View [396] provides street-level panoramic imagery captured in hundreds of cities on a global scale, allowing to audit neighborhood environments more easily. Rundle et al. [397] tested the feasibility of collecting data from Google Street View in 2008 to audit the environments of neighborhood in New York City (NYC). They found that measurements based on Google Street View exhibit higher levels of concordance in pedestrian safety, traffic and parking and infrastructure for active travel compared to field audit data in 2007. The result suggests the promising application of street views for auditing neighborhood environments in a more rapid way but with a lower cost.

Geotagged images collected from Google Street View and other websites have been leveraged to quantity urban perception using crowdsourcing methods, where human observers make a variety of perceptual inferences about images of places based on their prior knowledge and experiences. Salesses et al. [398] presented a method to measure the urban perception of safety, class and uniqueness in two US cities (Boston and NYC) and two Austrian cities (Linz and Salzburg) based on hundreds of geotagged images. They created a website to collect perception data by asking human observers to do pairwise comparison of two randomly selected images in response to questions, for example, on safety: "Which place looks safer?" (see Place Pulse, <http://pulse.media.mit.edu>). With the collected perception data, they calculated a  $Q$ -score for image  $i$  on question  $u$  by

$$Q_{i,u} = \frac{10}{3} \left( W_{i,u} + \frac{1}{n_i^w} \sum_{j_1=1}^{n_i^w} W_{j_1u} - \frac{1}{n_i^l} \sum_{j_2=1}^{n_i^l} W_{j_2u} + 1 \right), \quad (49)$$

where  $W_{i,u} = w_{i,u}/(w_{i,u} + l_{i,u} + t_{i,u})$  and  $L_{i,u} = l_{i,u}/(w_{i,u} + l_{i,u} + t_{i,u})$  are respectively the win (W) and loss (L) ratios of image  $i$  with  $n_i^w$  and  $n_i^l$  being respectively the total number of wins and losses. The  $Q$ -score takes the value in  $[0, 10]$  with  $Q = 10$  meaning the highest level of safety. They found that the range of perceptions elicited by images from Boston and NYC is wider, suggesting the two US cities are perceptually more unequal than Linz and Salzburg. Moreover, the spatial variation of urban perception helps explain violent crimes in NYC zones at zip-code resolution. Later, Ordóñez and Berg [399] predicted human perceptions of safety, uniqueness and wealth in urban places by applying classification and regression models to the Place Pulse dataset and another crowd-sourced dataset of street view images. They found that perceptual predictions are highly correlated with official crime and wealth statistics.

Novel computational tools have been used to predict urban perception from street images. Naik et al. [400] trained a scene understanding model (named *Streetscore*) based on data from an online survey with 7000 participants to predict the perceived safety of a streetscape using generic image features (see Figure 25). They used the Microsoft Trueskill algorithm [401] to convert ratings from the Place Pulse dataset to a  $Q$ -score for each image and then trained the  $\nu$ -support vector regression ( $\nu$ -SVR) model [402] using input feature vectors  $x$  and their corresponding labels  $y$  to predict the  $Q$ -score of each image. The goal of SVR [279, 280] is to approximate  $y$  by a regression function  $f(x)$ . Here,  $f(x) = (w \cdot w) + b$ . The key idea of  $\nu$ -SVR is to guarantee that the number of predictions with an error over  $\epsilon$  is

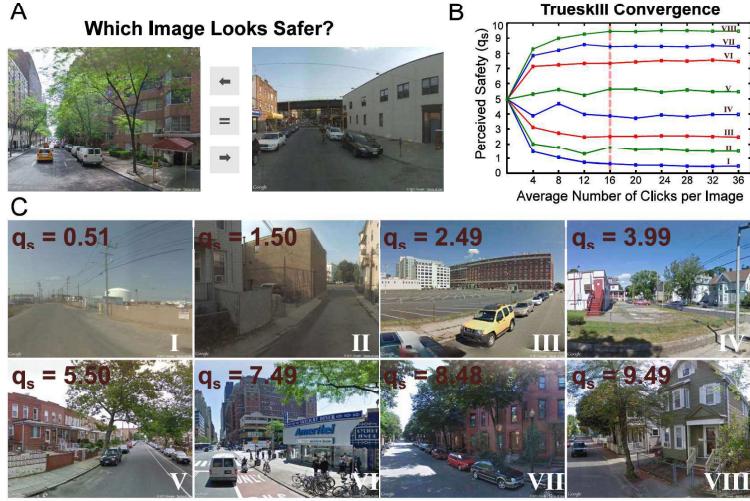


Figure 25: Perceived safety scores converted from pairwise image comparisons. (A) The pairwise image comparisons from a crowdsourced study, the Place Pulse [398]. (B) The Trueskill [401] converges to a stable  $Q$ -score of perceived safety after about 16 clicks. (C) The images are ranked by their  $Q$ -scores of perceived safety, which is between 0 and 10. Figure from [400].

less than  $v$  through minimizing the following function

$$\begin{aligned} \min \quad & \frac{1}{2} \|w\|^2 + C \cdot (v\epsilon + \frac{1}{K} \sum_{i=1}^K (\xi_i + \xi_i^*)) \\ \text{s.t.} \quad & \begin{cases} (w \cdot x + b) - y_i \leq \epsilon + \xi_i \\ y_i - (w \cdot x + b) \leq \epsilon + \xi_i^* \\ \epsilon \geq 0, \xi_i^* \geq 0 \end{cases}. \end{aligned} \quad (50)$$

The model including all features exhibits an accuracy of  $R^2 = 0.5676$  in predicting the  $Q$ -score of safety. After scoring about 1 million street view images using the *Streetscore*, they created high resolution perception maps for 21 US cities. However, Porzi et al. [403] argued that taking a “ground-truth” function in the regression model [398] or using the Trueskill scoring function [400] to predict scores of images based on users’ judgments will bring intrinsic bias. It is hard to explain the true distribution of users’ pairwise judgments by using one of many possible scoring functions. On the other hand, training an algorithm for prediction using a scoring function will leave no independent data for error assessments as the scoring function is constructed based on all users’ judgments. To address these issues, they developed a ranking framework that employs a convolutional neural network (CNN) to train algorithms directly on users’ ratings. Their CNN architecture uses a novel pooling layer to automatically discover mid-level visual patterns that have the strongest correlations with urban perception. Evaluations on the Place Pulse dataset demonstrates the advantages of their method on predicting perceived safety of images.

Yet, only using human judgments of images, it is still difficult to define some concepts in urban perceptions (e.g., beauty, happiness and quietness). To this end, Quercia et al. [404] explored visual assessments to better understand people’s visual perceptions and developed methods to automatically extract aesthetically informative features from city scenes. They built a crowdsourcing website (<http://urbangems.org>) to collect user ratings of London’s streets and then translated the ratings into quantitative measures of urban perception on three qualities: beautiful, quiet and happy. Next, they employed image processing techniques to determine visual cues that may cause the perception of a street. After checking the association with the three qualities, they found that the most positive visual cue is the amount of greenery. Arietta et al. [405] developed a method to automatically identify relations between a city’s visual appearance and its non-visual attributes. They spatially interpolated the data of (location, value of attribute) to obtain the attribute values and then detected discriminative visual elements of the attribute by building some SVMs [406]. Next, they trained attribute predictors by employing a nonlinear SVR [279, 280] to learn a set of weights over these elements. They found that visual elements are predictive to many city attributes including crime rates, population density and tree presence. The attribute predictors estimate the theft rate on average 33% more accurate than humans.

Computer vision methods have been increasingly used to learn deep features for scene recognition at the presence of the availability of large datasets. The introduction of deep convolutional neural networks (DCNNs) [20] has dramatically improved the performance in computer vision tasks, for instance, extracting urban perceptions from images. Seresinhe et al. [407] quantified the beauty of outdoor places by analyzing online crowdsourced ratings of over 200,000 images of Great Britain and features extracted from a scene-centric image dataset [408]. They found that places with natural features and man-made structures are considered more scenic. Then, they trained a neural network (named Scenic CNN) to predict the beauty of scenes for new places. The Scenic CNN can automatically identify natural and built scenic places such as the Big Ben in London. Dubey et al. [409] created a new global crowdsourced dataset (Place Pulse 2.0) consisting of over 1 million pairwise comparisons for more than 110 thousand images from 56 cities in 28 countries. Images are scored along perceptual attributes: safe, lively, boring, wealthy, depressing, and beautiful. They designed the Streetscore-CNN (SS-CNN) to predict the winner in the task of a pairwise comparison and then designed the Ranking SS-CNN (RSS-CNN) to better understand the fine-grained differences between image pairs. Experimental results show that RSS-CNN outperforms SS-CNN in predicting perceptual attributes. Moreover, models trained to predict one visual attribute can be used to predict other visual attributes with a fair accuracy.

Albert et al. [410] identified patterns in urban environments based on satellite imagery from Google Maps Static. They applied computer vision techniques based on DCNNs to classify images. Their results show good agreement with public benchmark data in green urban areas, water bodies, urban fabric, airports, etc. Moreover, they found that deep features of urban environments extracted from satellite imagery exhibit good performance on comparing neighborhoods across several cities, suggesting their method's ability of transfer learning. Tracewski et al. [411] investigated whether a deep learning network trained on scene characteristics can be used to classify volunteered photos for land cover characterization. They applied a simple post hoc weighting approach and a complex decision tree approach to extract land cover information from the network. They found that a general neural network without specific training can achieve modest levels of classification accuracy, suggesting the usage of well-validated methods without doing a long and costly training exercise. Lefévre et al. [412] explored the use of multiview imagery combining overhead and ground views on scene analysis. They suggested to integrate remote sensing, computer vision and machine learning for better urban observation.

Mobile phone (MP) data have also been used to visually profile cities and analyze urban lives. Based on GPS locations recorded by a smartphone app and self-rated happiness of over 20,000 subjects in the UK, Guillen et al. [413] explored the relationship between happiness and natural environments associated with GPS locations. They found that participants exposed to green environments or natural habitat types are significantly happier than those to urban environments. De Nadai et al. [414] explored whether safer looking neighborhoods are more lively in two major Italian cities (Rome and Milan). They defined metrics for human activity or liveliness in an area based on MP billing and operation data. Then, they employed DCNNs trained on the Place Pulse dataset to predict the scores of perceived safety based on streetscape imagery. Finally, they explored the relation between safety perception and liveliness using the spatially corrected ordinary least squares regressions. They found that neighborhoods perceived safer are more lively, while population demographics affect the perception. Moreover, street facing windows and greenery also contribute to safety perception. Recently, Harvey and Aultmanhall [415] reviewed approaches on measuring urban streetscapes for livability such as the uses of Internet-enabled surveys, streetscape images and social media.

Physical appearances of neighborhoods are not static, but changing over time. Naik et al. [416] introduced a computer vision method to understand physical dynamics of cities based on street views at different times. They collected over one million image cutouts for street blocks in 2007 and 2014 for five large US cities and matched the 2007 panel with the 2014 panel according to geographical locations of image cutouts. Then, they predicted the perceived safety of street view images using a variant of the Streetscore algorithm [400] and obtained Streetchange by calculating changes in streetscores of images from 2007 to 2014 (see Figure 26). After exploring the connection between Streetchange and socioeconomic changes, they found supportive evidences to the three classical theories of urban change: (i) both education and population density predict physical improvements in neighborhoods, supporting theories of human capital agglomeration; (ii) physical proximity to city centers and attractive neighborhoods predicts neighborhood improvement, supporting the *invasion theory* [417] that improvements in a neighborhood will spillover to adjacent areas; (iii) neighborhoods with better initial appearances have larger improvements, supporting the *tipping theory* [418] that nicer neighborhoods will get better.

Recently, data from other platforms have also been used to analyze neighborhood change and to infer socioeconomic levels. Glaeser et al. [419] explored the potential use of Yelp data in tracking neighborhoods change and

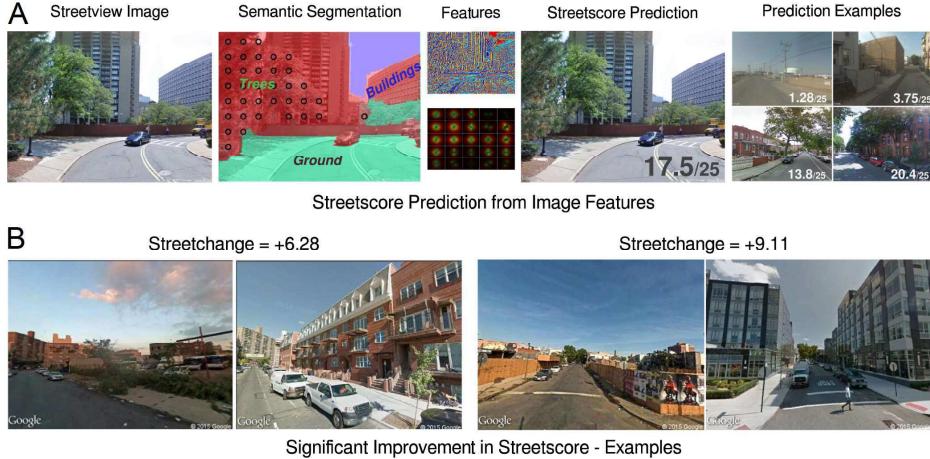


Figure 26: The computation of Streetchange. (A) Computing Streetscore using a regression model based on two image features (GIST and texton maps), which are extracted from pixels of four object categories (ground, buildings, trees, and sky) using semantic segmentation. (B) Computing Streetchange by the difference between the Streetscores of a pair of images captured in 2007 and 2014. Significantly positive Streetchange is usually associated with major construction. Figure after [416].

forecasting local economic activity. They found that the entry of Starbucks recorded by Yelp data is indicative of house price growth in the US, and gentrifying neighborhoods can attract more upscale establishments. Moreover, Streetchanges of perceived neighborhood quality are predictive of changes in local economy. Brelsford et al. [420] explored the topology and spatial evolution of neighborhoods based on a diverse set of detailed urban maps. They found that neighborhoods in developed cities fall into the same topological class while urban slums display different topological characteristics. Moreover, it is possible to build a street network in existing slums that can guarantee universal connectivity at minimal disruption and construction costs. Indeed, urban forms are predictive to socioeconomic status. Venerandi et al. [421] explored the relations between metrics of urban form and socioeconomic status extracted from five openly accessible datasets. They found that urban form can explain up to 70% of the variance in the official Index of Multiple Deprivation (IMD) of six major UK cities. Moreover, they observed some patterns of more deprived UK neighbourhoods such as higher population density, more regular street patterns, and more dead-end roads.

### 3.3.4. Urban computing for better lives

In the emerging interdisciplinary field named *urban computing* [422, 423], the unobtrusive and continuous improvement of urban lives have been increasingly studied in recent years. By leveraging data generated in cities, urban computing connects computer sciences with conventional city-related fields through technologies of urban sensing, data analytics and visualization in a recurrent process, with applications to social science, economy, urban planning, transportation, and so on. Recently, Zheng et al. [422] reviewed the general framework and key challenges of urban computing from a computer science perspective. They classified the applications into seven categories (i.e., urban planning, transportation, environment, public safety and security, energy, social, and economy) and summarized the typical technologies into four folds (i.e., urban sensing, urban data management, knowledge fusion across heterogeneous data, and urban data visualization). Later, Calabrese et al. [424] reviewed the techniques related to the use of mobile phone (MP) networks for urban sensing. They provided recommendations on datasets and techniques for specific applications. Regarding the improvement of urban lives, Glaeser et al. [425] described a variety of new urban data sources and illustrated how these data can be used to improve the quality of urban services.

Understanding what creates a better urban life is critical if anyone would like to make some improvements. Jacobs [426] suggested the urban physical environment as an essential factor for urban vitality. Specifically, the presence of pedestrians at all times of the day create life, while the elimination of pedestrian activity causes death. Therefore, the improvement of urban life in large cities is associated with the diversity of physical environments that require four essential conditions: mixed use, small blocks, buildings diverse, and building concentration. Recently, by collecting pedestrian activity through surveys and employing multilevel binomial models, Sung et al. [427] empirically verified

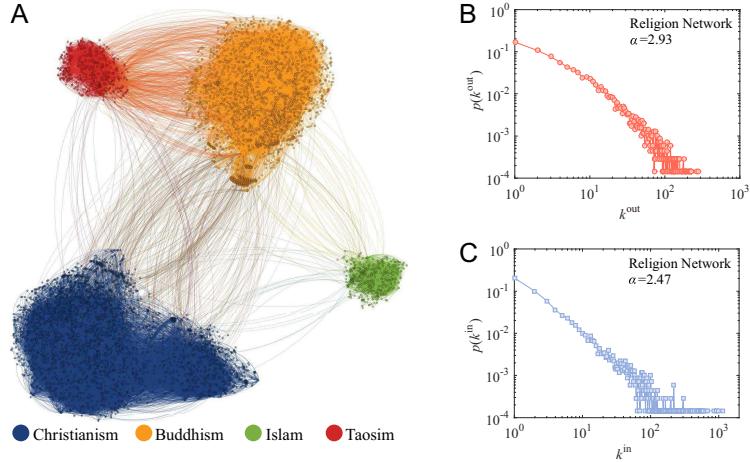


Figure 27: Structure of the religion network. (A) Structural layout of the network neglecting the directions of links, where blue, orange, green and red nodes denote Christians, Buddhists, Islamists and Taoists, respectively. (B) The out-degree distribution in a log-log plot. (C) The in-degree distribution in a log-log plot. Figure from [433].

the role of Jacobs's four conditions in urban diversity in Seoul, South Korea. De Nadai et al. [428] later verified the necessity of Jacobs's four conditions for the promotion of urban lives in six Italian cities. They used a proxy for urban vitality extracted from MP records, land usage mapped from satellite images, and socio-demographic information collected from national census and the Open Street Map project. Their work provides a new way to test traditional urban theories in fine-grained details.

Recent literature have studied urban segregation of people with different socioeconomic status. Using geotagged tweets in Louisville, Shelton et al. [429] developed an approach to study intra-neighborhood segregation, mobility and inequality. After analyzing the everyday activity spaces of different groups, they proposed to understand Louisvillian neighborhoods by the fluid, porous and actively produced. After exploring the socio-spatial segregation in Beijing, Wang et al. [430] found significant differences between residents inside and outside the so-called privileged enclaves in the usages of time. They suggested scientists should pay more attention to how different social groups actual use urban spaces and spend their time in terms of everyday activities. Yip et al. [431] analyzed the mobility patterns of people in Hong Kong that are tracked by a mobile phone app. They found that the interactions of people with other income groups are limited. Rich people tend to move to rich neighbourhoods, while poorer people move to poorer neighbourhoods. Louf and Barthélémy [432] provided a direct definition of residential segregation and showed that richer class in high density zones is over represented. In particular, they suggested density as a relevant factor for understanding urban income structure and explaining differences observed in cities.

Data from social networks have been used to study religious segregation and urban indigenization. Hu et al. [433] quantified religious segregation by analyzing religious social network based on Weibo. They found that the religious network is highly segregative (see Figure 27), and the extent of religious segregation is higher than racial segregation. In addition, 46.7% of cross-religion connections are probably related to charitable issues, suggesting the role of charitable activities in promoting cross-religion communications. Yang et al. [142] identified the distinct mobility patterns of natives and non-natives in five large cities in China by analyzing about 1.37 million check-ins. They found that the distribution of location visiting frequencies is relatively homogeneous for natives as they usually check in repeatedly at locations of personal importance. By contrast, the distribution is more heterogeneous for non-natives as they tend to visit popular locations. With this insight, Yang et al. [142] proposed a so-called indigenization coefficient to estimate the likelihood of an individual to be a native or to what extent an individual behaves like a native, which is based solely on check-in behaviors. Such method can be applied in estimating the time required for non-natives to behave the same as natives, as well as in enhancing the prediction accuracy of human mobility (i.e., the next check-in location).

The Schelling model [434] is widely used to explain the emergence of racial segregation. A small preference to alike neighbors at the individual level can lead to large-scale segregation at the collective level through neighborhood

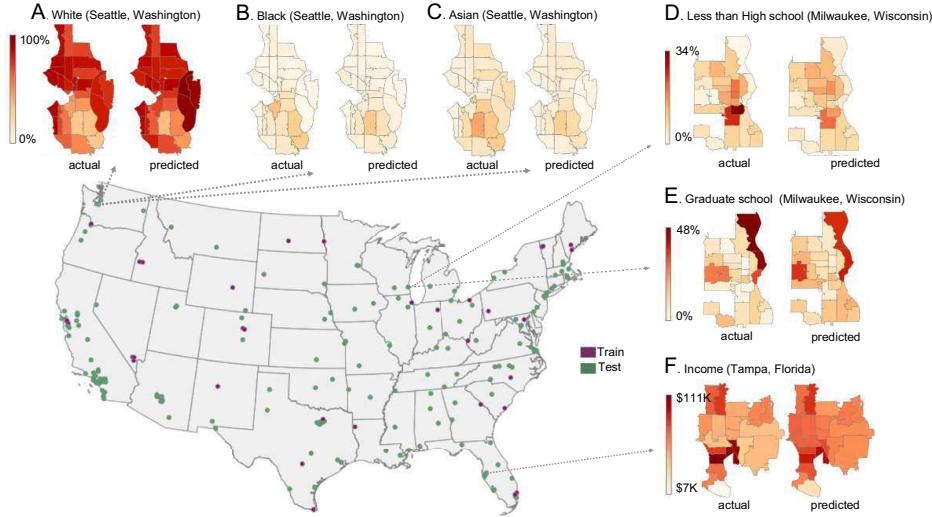


Figure 28: Estimations of socioeconomic status from car attributes. The model was trained in counties (shown in purple on the map) and used to estimate demographic variables at the zip code level for all cities (shown in green). Maps of actual versus predicted values are shown for the percentages of (A-C) Black, Asian, and White people, (D) people with less than a high school degree, (E) people with graduate degrees, as well as for (F) the median household income. Figure after [437].

tipping. Sahasranaman et al. [435] studied the dynamics of transformation from segregation to mixed wealth cities using a variation of the Schelling model, in which the movement of agents to neighborhoods should satisfy the threshold condition: the wealth of agents is not lesser than the wealth of the threshold proportion of their neighbors in the new neighborhoods. They found that wealth-based segregation occurs and persists, however, the dynamics can yield a persistent mixed wealth distribution in the tolerance condition that small proportions of disallowed moves (where threshold condition is not satisfied) are introduced. The results suggest to enable a small fraction of disallowed moves to drive the transformation to mixed wealth cities. Later, Sahasranaman et al. [436] extended the above model and studied the long-term patterns of neighborhood economic status. They found that very poor and very rich neighborhoods tend to retain economic status than middle-wealth neighborhoods.

Deep-learning-based computer vision techniques have been applied to analyze digital imagery, which provides a faster and cheaper alternative of community survey. Gebru et al. [437] proposed a method to estimate socioeconomic trends from 50 million street view images in 200 US cities (see Figure 28). They automatically detected 22 million distinct vehicles from images using the object recognition algorithm [438] and then deployed CNNs [47] to determine features of vehicles and classify each vehicle into one of the 2,657 fine-grained categories. Using the resulting data, they estimated race and education levels by training a logistic regression model and estimated income and voter preferences by employing a ridge regression model. Compared to the American Community Survey, their demographic estimates exhibit satisfied accuracy at the city level (e.g.,  $r = 0.82$  for the estimates of median household income,  $r = 0.87$  for the percentage of Asians, and  $r = 0.70$  for the percentage of people with a graduate degree). The method can also provide a good accuracy at a more fine-grained zip code resolution, for example, the estimation of the percentage of Asians yields a correlation  $r = 0.77$  at zip code resolution for Seattle. In addition, the method can accurately estimate voter preference. For example, a city tends to vote for a Democrat (88% chance) if encountering more sedans than pickup trucks during a drive through the city.

The demand for facilities (such as hospitals, airports and malls) increases as the development of cities. Locations of facilities are ideally determined by the necessities of people who live nearby. Theoretical derivations starting from assumptions of least cost [439] and minimum time [440] have suggested that the optimal density of facilities  $D(r)$  at certain position  $r$  scales as a power law ( $\alpha = 2/3$ ) with the density of relevant population  $\rho(r)$ . After analyzing the distribution of over 400 nongovernmental and service establishments in the US, Stephan [441] found that  $\alpha$  is more close to 1 than to the widely observed  $2/3$  for governmental establishments. Gastner and Newman [442] found a slope of 0.66 in the log-log plot of  $D(r)$  versus  $\rho(r)$  in the US and presented an analytic solution based on density-dependent

map projections. Recently, Um et al. [443] explored the scaling  $D \sim \rho^\alpha$  and found that  $\alpha$  depends on facility types. By proposing a microscopically mechanism model, they demonstrated that public facilities driven by social opportunity cost have an exponent  $\alpha \approx 2/3$ , whereas private facilities driven by profit have an exponent  $\alpha \approx 1$ . The distributions of the optimal positions of public or private facilities on real US map predicted by their model agree well with the empirical data.

Fine-grained data describing business activities have been used to identify the optimal locations for new retail stores and improve the layout of amenities (e.g., schools, restaurants, cafes, and libraries) in urban areas. Karamshuk et al. [444] predicted the popularity of retail stores in NYC using machine learning approaches based on check-in data collected from Foursquare. They tested the predictive power of various features extracted from the check-in data including geographic features and mobility features, finding that the strongest indicators of popularity are the presence of user attractors and retail stores of the same type. Recently, Hidalgo and Castañer [445] studied the neighborhood-scale agglomerations of amenities by building a network of amenities, named amenity space, based on the precise locations of millions of amenities across 47 US cities. They introduced a clustering algorithm to identify neighborhood-scale agglomerations and mapped the amenity space by connecting pairs of amenities that are more likely to co-locate in the same neighborhood (see Ref. [77] for a similar method). Based on the amenity space, they further built a recommender system [446] to recommend missing amenities in the neighborhood, given its current pattern of specialization. Their method provides new avenues for the optimal layout of facilities within cities.

The analysis of bank card transactions provides a new way to estimate regional socioeconomic status and improve urban spatial equity. Louail et al. [447] analyzed a database of card transactions in two Spanish cities (Madrid and Barcelona) and then proposed a bottom-up approach to redistribute money flows for equality situations through redirecting a limited fraction of individual shopping trips. They constructed the “individual-business” bipartite spatial network, where the edges correspond to card transactions. Then, they performed the rewiring of individual transactions by redirecting them to the same business category located in different neighborhoods, with the goal to re-balance the commercial income among neighborhoods and with the preservation of human mobility properties (see Ref. [447] for details). They found that reassigning only 5% of individual transactions can reduce more than 80% spatial inequality between neighbourhoods and can even improve other sustainability indicators like total distance traveled and spatial mixing. Their work illustrates an excellent implementation of crowdsourcing the “Robin Hood effect” [448], a process through which capital is redistributed to reduce inequality.

## 4. Individual socioeconomic status and attributes

### 4.1. Individual socioeconomic level

Socioeconomic level is an indicator typically defined as a combination of income related variables to characterize an individual’s social and economic status. As such, individual socioeconomic level serves as an indication of the purchasing power, and thus it is important to the design and evaluation of social policies. In this subsection, we will introduce literature on how to infer individual socioeconomic levels from the ownership and usage of mobile phones (MPs), social media (SM) data, bank transitions, human mobility patterns and individual behavioral traits.

#### 4.1.1. Mobile phone and credit card usage

The computation of socioeconomic indices faces some challenges such as the large expenses in acquiring a whole country’s data, the long-time delay of the census data, and the lack of high-quality data in developing economies. Thanks to the development of information and communication technology, MPs are now becoming ubiquitous and frequently used even in the world’s poorest countries and regions [449]. MPs serve as important sensors of human activities, which should be closely related to individual socioeconomic status. Using behavioral variables extracted from MP data, Soto et al. [450] predicted individual socioeconomic level by employing models constructed with the support vector machines and random forests. Using only 38 features, their model can achieve a classification accuracy up to 80% in determining socioeconomic levels of about 0.5 million citizens. With promising applications, their method can be a cost-effective complement to traditional socioeconomic estimation techniques.

The ownership and usage of MPs in the past decade are not uniform across populations and continents. Blumenstock and Eagle [451] provided a quantitative perspective on the socioeconomic structure of individual MP usage in

Rwanda between 2005 and 2009. After analyzing data collected through interviews and merged with MP call histories, they found that MP owners are considerably wealthier, better educated and more predominantly male. Most notably, they found some differences in MP adoption between the relatively poor and rich. For example, richer people have a larger number of calls, a greater total length of calls, more MP-using days, a larger number of contacts, and so on. Their results demonstrate that MPs are owned and used by the privileged strata of Rwanda society (see also Ref. [452]). Wesolowski et al. [453] analyzed a survey of MP ownership and usage across Kenya in 2009 to understand the social and geographical heterogeneity of MP usage patterns. They found distinct regional, gender-related and socioeconomic variations, with particularly low ownership among rural communities and poor people. In particular, there is a nonlinear relationship between MP ownership and sharing behavior across counties, where MP sharing practices are extremely common in rural areas.

Novel methods have been proposed to infer individual socioeconomic levels from MP data. Blumenstock [454] explored the extent to which MP data can be used to predict an individual's socioeconomic status by analyzing the combined data of MP records and phone-based interviews in Rwanda during 2009-2010. They found significant correlations between asset ownership and some MP-derived measures including phone usage, social network structure and geographic mobility. The first principal component of 97 metrics of MP usages can explain 34.63% of the variance in the asset categories. Moreover, simple classification methods are able to predict the fixed household characteristics and whether the respondent owns assets. The prediction accuracy for television ownership is over 85%. Agarwal et al. [455] analyzed 350 million MP call logs and found that phone-based features have significant predictive power for an individual's financial risk with an accuracy of about 65%. Recently, deep learning technologies are also used to classify individual socioeconomic status based on large-scale MP datasets. Sundsøy et al. [456] explored how socioeconomic levels can be accurately classified by implementing a multi-layer feed-forward deep learning architecture [20] without any manual operations on feature selection. The new model using location traces as the sole input achieves an average AUC value about 0.77 in separating individuals into high and low socioeconomic status.

MPs can generate rich data about financial histories of mobile money. For example, airtime credit is the money on MP devices, which can be used for purchase (e.g., calls, texts and data) and be transferred to others. Based on the history of airtime credit purchases and MP communications in 2012, Gutierrez et al. [59] estimated the relative income of individuals and the diversity of income for fine-grained regions in Côte d'Ivoire. They quantified individual purchasing behavior by calculating the variation in the purchase amounts through the coefficient of variation (CV):

$$CV = \sigma/\mu, \quad (51)$$

where  $\sigma$  is the standard deviation, and  $\mu$  is the mean value. They found that some people make few big purchases, while others make many small purchases. Thereby, they hypothesized that the frequency and size of purchases are correlated with individual income as the poorer may not buy lots at once due to the lack of enough airtime credit. After analyzing the social network built on the MP communications, they found a certain homophily in terms of purchase averages, where people belonging to the same community tend to have the same amount of average purchase. Recently, behavioral patterns revealed by MP usages have been used to predict default among borrowers. For example, Björkegren and Grissen [457] proposed a method to predict the likelihood of repayment using behavioral features derived from MP transaction records in a Caribbean country. The method achieves an AUC of 0.76 and 0.77 respectively for banked and unbanked consumers without formal financial histories.

Purchasing behaviors documented by bank accounts and credit cards are also predictive to individual socioeconomic status. By analyzing bank administrative data and survey data, Prina [458] found that the access to free savings accounts can generate welfare effects. After randomly giving free access of bank accounts to many female household heads in Nepal, they found that physical proximity and zero fees lead to high take-up and usage rates of savings accounts, resulting in the overall improvement of financial situation for low-income households. These results suggest that the access to formal financial services can lead to household improvements. Dong et al. [459] analyzed millions of individual credit card transactions in two low-to-middle income countries. They found that patterns of purchase activity are strongly correlated with socioeconomic status. They defined two measures of purchase diversity at the district-level (the outgoing purchase diversity and the incoming purchase diversity) by averaging the purchase diversity defined by the Shannon entropy (see Ref. [248] for details) over corresponding individuals. They found a positive correlation between both purchase diversities and socioeconomic status, e.g.,  $r = 0.77$  for the European country and  $r = 0.53$  for the Latin American country in case of outgoing purchase diversity.

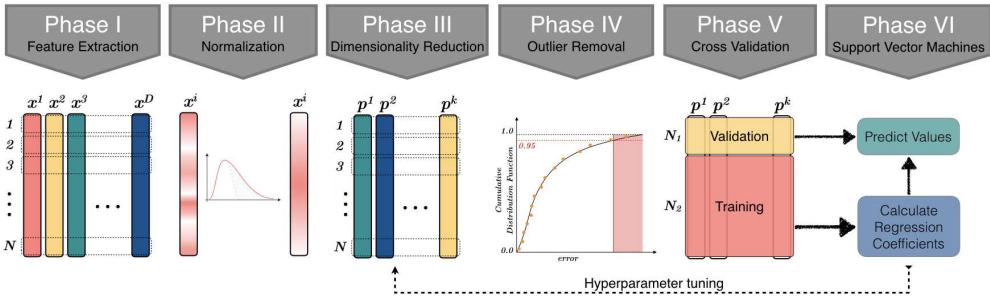


Figure 29: The modular machine learning workflow. The individual financial transactions are used in the predictive modeling of official socioeconomic indicators. The support vector machines (SVM) is employed based on the reduced features space using the principal component analysis (PCA). Figure from [460].

Hashemian et al. [460] connected individual financial transactions with the socioeconomic description of regions by analyzing a dataset of bank card transactions in Spain. They built a feature space after extracting 33 microeconomics indicators and predicted official socioeconomic indices by employing a modular machine learning workflow (see Figure 29). Their model performs decently, e.g., in predicting regional GDP per capita with an accuracy of 0.729 using only seven principal components of the reduced features space. Sobolevsky et al. [461] analyzed the same Spanish transaction data and found a clear correlation between individual spending behaviors and six official socioeconomic indicators. They created a feature space consisting of 35 microeconomics indicators derived from the bank card transactions. They further applied the principal component analysis (PCA) [462] to build a reduced feature space, based on which they used a generalized linear model (GLM) [155] to predict socioeconomic indices. Their model well predicts all the considered quantities. The average accuracy on the training set is about 0.70 for the province-level GDP prediction. By analyzing credit card purchasing data, Clemente et al. [463] found that male and young adults are of higher probability to use credit card than female and old adult. Moreover, the median credit card record expenditure is correlated with the average monthly wage at the district level.

#### 4.1.2. Social profile and network structure

Individual socioeconomic status can be inferred from social media (SM) data and the network structure of online social networks. For example, there are wide discussions on the embeddedness of economic activities in social networks [247]. On the other hand, users' behaviours on social network are affected by their socioeconomic status and health consciousness. For example, the non-adaptive coping responses to health-related messages in the mass media are negatively correlated with socioeconomic status [464]. Wangberg et al. [465] explored the relations between individual Internet use and socioeconomic status based on two survey datasets. They found that Internet is a plausible mediator between subjective health and socioeconomic variables.

Messages posted to SM have been used to estimate users' social class and socioeconomic status. Filho et al. [466] proposed a method to predict a user's social class based on interactions on Foursquare and messages on Twitter with a hypothesis that richer users usually visit wealthier places. In their method, each neighbourhood is assigned with a social class according to the income. They mapped the coordinates of Foursquare interactions and tweets to neighbourhoods and assigned every visited place with a social class. A user's social class is estimated by the most frequently visited class. Further, they employed classification models to predict users' social classes using textual features extracted from tweets. The average F1 value of their models varies from 0.57 to 0.73, depending on the classification segments. Based on a large dataset of Twitter users annotated with income, Preoziuc-Pietro et al. [467] built a model to predict income using user profile, psycho-demographic, emotion and shallow textual features. They found that high earnings are indicated by intelligence, high education, male and old age. Low-income users express more disgust emotions and sadness, have high posting rate, use more swear words, and post more URLs. In contrast, high-income users express more fear and anger, have more retweets and talk more about justice and politics. They proposed nonlinear models that can predict user income with an accuracy of 0.633 using a combination of all features.

Based on UK Twitter users' profiles and tweets, Lampous et al. [468] proposed a method to classify users into the upper, middle or lower socioeconomic status. They mapped each user to a socioeconomic status by utilizing the

standard occupation classification hierarchy [469]. They extracted five categories of features from tweets: behavior, impact, profile, tweets, and topics. They performed the classification by employing a nonlinear learning approach, in which a composite Gaussian process (GP) is used. Given the input  $\mathbf{x} \in \mathcal{R}^d$ , the aim of the GP method is to learn a function  $f : \mathcal{R}^d \rightarrow \mathcal{R}$ :

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')), \quad (52)$$

where  $m(\cdot)$  is the mean function. The covariance kernel  $k(\cdot, \cdot)$  for feature category is given by

$$k(\mathbf{x}, \mathbf{x}') = \left( \sum_{n=1}^C k_{SE}(\mathbf{c}_n, \mathbf{c}'_n) \right) + k_N(\mathbf{x}, \mathbf{x}'), \quad (53)$$

where  $\mathbf{x} = \{\mathbf{c}_1, \dots, \mathbf{c}_C\}$  is the input of  $C$  feature categories, and  $k_N(\mathbf{x}, \mathbf{x}') = \theta_N^2 \times \delta(\mathbf{x}, \mathbf{x}')$  models noise with  $\delta$  being the Kronecker delta function. The  $k_{SE}(\mathbf{x}, \mathbf{x}')$  defines the squared exponential kernel, which is given by

$$k_{SE}(\mathbf{x}, \mathbf{x}') = \theta^2 \exp(-\|\mathbf{x} - \mathbf{x}'\|_2^2 / (2\mathcal{L}^2)). \quad (54)$$

The GP model yields accuracies of 0.7509 and 0.8205 respectively for 3-way and binary classification scenarios.

The structure of mobile phones (MP) communication networks has been used to predict individual socioeconomic status. Leo et al. [470] analyzed a coupled dataset of MP call detailed records (CDRs) and bank credit information of individuals living in Mexico. They found that the identified socioeconomic classes from the structure and evolution of the communication network are strongly correlated with typical consumption patterns. Moreover, they found positive correlations between people regarding their economic status and further confirmed the social stratification in social structure. Fixman et al. [471] analyzed CDRs and account balances for over 10 million clients of a Mexican bank and found a strong socioeconomic homophily in Mexico. Users linked in the MP communication network are more likely to have similar income. Further, they proposed a Bayesian approach to predict individual income, where individuals are distinguished into either low income group ( $H_1$ ) or high-income group ( $H_2$ ) according to their income ( $g_s$ ). For user  $q^j$ , the amount of outgoing calls  $a_i^j$  to the category  $H_i$  is calculated. The Beta distribution  $B^j$  for the probability of belonging to a given category is defined based on  $a_i^j$ :

$$B^j(x; \alpha^j, \beta^j) = \frac{1}{B(\alpha^j, \beta^j)} x^{\alpha^j-1} \cdot (1-x)^{\beta^j-1}, \quad (55)$$

where  $\alpha^j = \alpha_1^j + 1$  and  $\beta^j = \alpha_2^j + 1$  are parameters of the Beta distribution

$$B(\alpha, \beta) = \frac{\Gamma(\alpha) \cdot \Gamma(\beta)}{\Gamma(\alpha + \beta)}, \quad (56)$$

where  $\Gamma(\alpha)$  is the gamma function. After obtaining the Beta distribution for the probability of belonging to the high-income category, the lowest five percentile  $p_{lower}$  for this probability is found. The user's income category is set to  $H_1$  if  $p_{lower}$  is above a given threshold  $\tau$  and  $H_2$  if otherwise. Their method achieves an accuracy of AUC=0.71 with  $\tau = 0.4$  in classifying users into low- and high-income groups.

For a Latin American country's whole population, Luo et al. [472] built a giant connected social network among 107 million users based on data about MPs and residential communications, and estimated individuals' financial status based on the combined credit limit of their credit cards. They found that people in the top economic class have higher diversity in communicating with equally affluent people and in connecting remote locations (see Figures 30A-D), suggesting the significant different communication patterns across economic classes. They further measured an individual's centrality in the network using the so-called collective influence (CI) [473], in addition to degree, PageRank [63] and k-shell index [474]. The k-shell index  $k_s$  of a node is the location of the shell obtained by iteratively pruning all nodes with degree  $k < k_s$  (see Figure 30E). The CI index of an arbitrary node  $i$  can be obtained by the optimal percolation theory [473], say

$$CI = (k_i - 1) \sum_{j \in \partial \text{Ball}(i, l)} (k_j - 1), \quad (57)$$

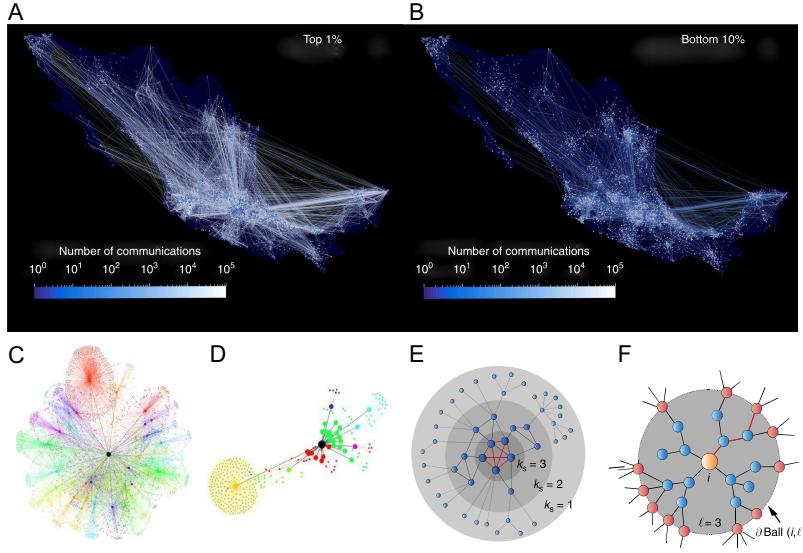


Figure 30: Communication networks of the population (A) in the top 1% and (B) in the bottom 10% of credit limit classes. Nodes are communities with the size denoting the number of bank clients inside each community. The colour and thickness of the edges reflects the number of communications. Examples of the ego-network for an individual (C) in the top 1% wealthy class and (D) in the bottom 10% class. (E) Illustration of the k-shell network decomposition. (F) Illustration of the calculation of the CI index. The CI Ball( $i, l$ ) of radius  $l = 3$  around the central node  $i$  (yellow) is the set of nodes inside the sphere, and  $\partial\text{Ball}$  is the set of nodes on the boundary (brown). The value of CI is the degree-minus-one of the central node times the sum of the degree-minus-one of the nodes at the boundary of the sphere of influence. Figure from [472].

where  $k_i$  is the degree of the node  $i$ ,  $\text{Ball}(i, l)$  is the set of nodes within radius  $l$  centred at node  $i$ , and  $\partial\text{Ball}(i, l)$  is the set of nodes at the radius  $l$  on the boundary (see Figure 30F). They found that individual economic status is highly correlated with centralities in the social network. The correlation for k-shell is  $R^2 = 0.96$  and for CI ( $l = 2$ ) is  $R^2 = 0.93$ . The correlation for CI can be increased by including the ages of individuals. Formally, they defined a new index

$$\text{ANC} = \alpha \text{Age} + (1 - \alpha) \text{CI}. \quad (58)$$

ANC exhibits a hight correlation  $R^2 = 0.99$  when  $\alpha = 0.5$  and  $l = 2$ . They further quantified the diversity of an individual's links by  $\text{DR} = W_{\text{out}}/W_{\text{in}}$  [249], where  $W_{\text{out}}$  and  $W_{\text{in}}$  are the total communications with this individual outside and inside his/her own community, respectively. They found that an age-diversity composite

$$\text{ADC} = \alpha \text{Age} + (1 - \alpha) \text{DR} \quad (59)$$

correlates well with individual economic status ( $R^2 = 0.96$ ) when  $\alpha = 0.5$ . The ability of communicating with individuals outside one's local tightly-knit social community and positioning oneself at network locations of high CI is suggestive to a high socioeconomic level.

By analyzing CDRs and income group of surveyed individuals, Jahani et al. [475] found that the structural diversity of ego networks exhibits a relatively strong correlation with the income of individuals. They built a communication network based on the CDRs, where the weights of links are the total number of calls between two connected individuals. Then, they calculated the structural diversity from the view of ego networks [476]. In particular, they measured the diversity of the alters (i.e., neighbors of the ego) by defining the weighted structural novelty, as

$$M_i = \frac{\sum_{j \in N(i)} (1 - \sum_{q \in N(i) \cap N(j)} p_{iq} p_{qj})}{|N(i)|}, \quad (60)$$

where  $i \neq j \neq q$ ,  $N(i)$  is the set of ego  $i$ 's neighbors, and  $p_{ij}$  is the proportion of time that ego  $i$  spent on its neighbor  $j$ . They found a positive correlation of the structural diversity on income after controlling education, occupation, age and gender.

Data from online gaming platforms have also been used to explore the relations between individuals' positions in social networks and their economic outputs. Xie et al. [477] analyzed a database recorded by 124 servers of a popular online role-playing game in China. They found that the position diversity of individuals in the game is positively correlated with their economic output and social status. For a given friendship network, a dependence network is built by removing the insignificant edges. In the dependence network, they identified 13 directed triadic motifs [478], within which 30 distinct motif positions [479] are located. Accordingly, they obtained the motif position ratio profile  $p_i = (p_{i,1}, \dots, p_{i,30})$  for an arbitrary individual  $i$ . For individual  $i$ , the individual position diversity  $d_i$  is defined by the Shannon entropy,

$$d_i = - \sum_{j=1}^{30} p_{i,j} \ln(p_{i,j}). \quad (61)$$

They found that  $d_i$  is highly correlated ( $r = 0.63$ ) with individual economic output. Further, they applied the  $k$ -means algorithm [480] to cluster individuals based on their position ratio profiles  $Z_i$ , where  $Z_i$  is the z-score of  $p_i$  (see Ref. [477] for details). The cluster centroid locations  $P_k = (P_{k,1}, \dots, P_{k,30})$  can be obtained for an arbitrary class  $C_k$  after measuring the closeness between position ratio profiles  $Z_i$ . The cluster position diversity of individual  $i$  is defined as  $D_i = -\sum_{j=1}^{30} P_{i,j} \ln(P_{i,j})$ . They found that economic outputs of classes increase with  $D_i$ . This work demonstrates that the structure of social network is predictive to individual economic outputs even in virtual world.

#### 4.1.3. Human mobility pattern

The relations between individual socioeconomic status and human mobility patterns have been explored in some scenarios based on a variety of new data resources [55, 481]. The relations are in two aspects. Socioeconomic status and demographic information have effects on mobility patterns of individuals [482]. For example, Carlsson-Kanyama and Linden [483] analyzed national travel survey data in Sweden. They found that middle-aged rich persons travel much farther, while low-income persons in general do not travel extensively. Propper et al. [484] analyzed hospital episode statistics in England and found that individuals in higher deprived wards travel less far for hospital admission. Fan et al. [482] revealed the correlation between social proximity and mobility similarity based on an LBSN dataset. On the other hand, large-scale mobility datasets (generated by MPs, digital cards, online platforms, and so on) have been used to quantify human mobility patterns at fine spatio-temporal scales, based on which individual socioeconomic status can be predicted. Lotero et al. [485] explored the role of socioeconomic differences in urban mobility from a multiplex perspective based on the origin-destination surveys carried out in two Colombian cities. Each city is represented by six multiplex networks, where each one represents the trips of individuals with a specific socioeconomic status (SES), from SES1 (low) to SES6 (high). In each multiplex network, layers correspond to different transportation modes (e.g., pedestrian and public transport), and layers are merged by subsequently adding the one representing the mostly used transportation mode. This process produces a mobility multiplex network (MMN) for each SES. Some structural measures of multiplex networks [486] are used to quantify the mobility pattern of each SES, and two overlap measures are defined to quantify the tendency of individuals with the same SES to use different transportation. They found that individuals belonging to SES1 display the smallest clustering coefficient and tend to avoid overlapping. Moreover, the poorest (SES1 and SES2) covers larger urban areas in a sparse way by using few and cheap transportation modes, the meddle (SES3 and SES4) covers most urban zones, and the elite (SES5 and SES6) covers smaller urban areas in a dense way by selecting costly transportation modes.

Based on the same Colombian survey data, Lotero et al. [487] further explored the temporal dependence of the trips performed by individuals with different SES. They found that the early-morning peak time delays (says rich do not rise early) and the midday peak becomes smoother as wealth increases. The strength of trips is more geographical dispersive for middle class but more localized for richest class (see Figure 31). Moreover, the efficiency of urban mobility (defined by the ratio of the average distance traveled to the average time spent per trip) increases with the socioeconomic level. Carra et al. [488] explored the relations between individual commuting distance and income based on datasets of national household surveys in Denmark, the UK and the US. Empirical results for Denmark and the UK confirm the prediction of basic equilibrium models [489] that individuals with a higher income have longer average commuting distances within a single city. The distribution of individual commuting distance across different countries is broad with a slow decaying tail that can be fitted by a power law with exponent  $\gamma \approx 3$ . Further, they proposed a new closest opportunity model for job searching process that can well predict the average commuting distance using individual income.

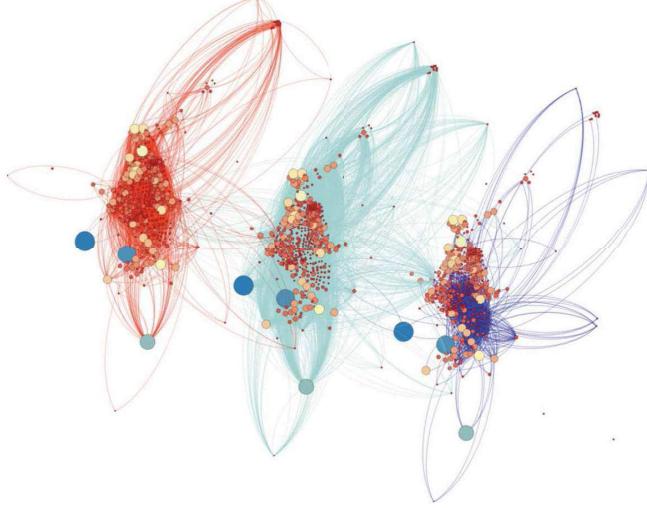


Figure 31: Three mobility networks of the Colombian city Medellín corresponding to three different socioeconomic status (SES). Networks from left to right correspond to SES1 (poorest class), SES3 (middle class) and SES6 (richest class), respectively. Nodes represent the origin and destination zones according to the origin-destination surveys. Sizes of nodes show the commuting strengths with colours representing their degrees. Figure from [487].

CDRs are increasingly used in human mobility analysis, which provides a chance to estimate individual socioeconomic status. Frias-Martinez et al. [490] analyzed CDRs in a Latin American country and found that population with higher socioeconomic levels have larger mobility ranges compared to population with lower socioeconomic levels. In particular, socioeconomic level is highly correlated with six human mobility variables including the traveled distance and radius of gyration [55]. The latter is defined as

$$r_g(t) = \sqrt{\frac{1}{n(t)} \sum_{i=1}^n (r_i - r_{cm})^2}, \quad (62)$$

where  $r_i$  with  $i = \{1, \dots, n(t)\}$  is the vector of phone tower  $i$ 's coordinates, and  $r_{cm} = \sum_{i=1}^n r_i / n(t)$  is the vector of the mobility trajectory's center of mass. There is a high correlation ( $r = 0.58$ ) between the number of used phone towers and the socioeconomic level. Further, they proposed a model that can estimate the tower-level socioeconomic status with an adjusted  $R^2 = 0.72$ . Later, Frias-Martinez et al. [491] analyzed a large-scale dataset of CDRs and suggested to predict future values of socioeconomic indicators based on human behavioral patterns. Using the multivariate regression analysis, they found that mobility variables perform better than consumption variables on predicting future socioeconomic indicators. Moreover, multivariate time-series models can yield high accuracy, for example, the training  $R^2$  value is about 0.68 in predicting the total assets.

Based on CDRs of about 20 million users in France, Pappalardo et al. [492] explored the relations between human mobility patterns and socioeconomic development. For each individual, they extracted a measure of mobility volume and a measure of mobility diversity from the CDRs. For individual  $i$ , the mobility volume (MV) is defined by the radius of gyration  $r_g(i)$ , and the mobility diversity (MD) [493] is measured by the Shannon entropy,

$$MD(i) = -\frac{\sum_{e \in E} p(e) \log p(e)}{\log N}, \quad (63)$$

where  $e = (a, b)$  stands for a trip from the origin base station  $a$  to the destination base station  $b$ ,  $E$  is the set of origin-destination pairs with size  $N$ , and  $p(e)$  is the probability of a movement along  $e$ . After aggregating the measures at the municipality level, they found that MD exhibits the strongest correlation with socioeconomic indicators (Pearson coefficient  $r = 0.49$  for per capita income,  $r = 0.49$  for primary education rate,  $r = -0.43$  for deprivation index, and  $r = -0.17$  for unemployment rate), showing that individuals living in developed municipalities have a high mobility

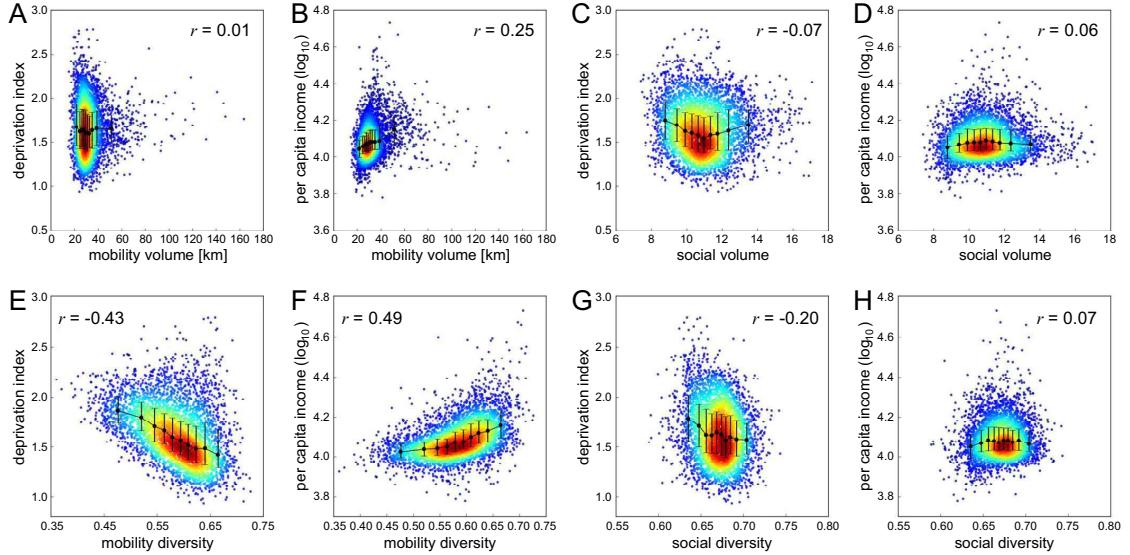


Figure 32: Relations between aggregated diversity measures and socioeconomic indicators. Mobility volume versus (A) deprivation index and (B) per capita income. Social volume versus (C) deprivation index and (D) per capita income. Mobility diversity versus (E) deprivation index and (F) per capita income. Social diversity versus (G) deprivation index and (H) per capita income. Black circles and error bars correspond to the mean and standard deviation for ten equal-sized municipalities. The  $p$ -value of the Pearson correlation coefficient  $r$  is under 0.001 in all panels. Figure after [494].

diversity. Moreover, there is a negative correlation ( $r = -0.38$ ) between MD and MV, suggesting that people living in less developed municipalities have to travel in search of activities out of their municipalities. Pappalardo et al. [494] further added two measures into consideration, namely, the social volume (SV) [18] and the social diversity (SD) [248]. They found that MD exhibits a higher correlation with socioeconomic indicators than SV, SD and MV (see Figure 32). Moreover, MD adds a predictive power to models that use social and mobility measures in the prediction of socioeconomic indicators.

Similarly, Florez et al. [495] extracted urban mobility patterns from CDRs covering 1.5 million users and studied many characteristics of the commuting network. They found a significantly positive correlation ( $r = 0.47$ ) between the income rank of a group and its commuting trip diversity [494]. Moreover, poor people travel longer to work and spend more time in commuting. Yang et al. [496] derived six mobility indicators from large-scale MP data and proposed a data fusion approach to approximate the aggregated socioeconomic status of MP users. They found that richer groups tend to travel longer in Boston but shorter in Singapore, however, different socioeconomic classes in both cities exhibit similar diversity of individual travel and activity patterns. Hong et al. [497] introduced a LDA-based topic modeling framework [263] to estimate socioeconomic levels based on MP data. They employed LDA to extract latent recurring patterns of population behaviors (topics) from individual behaviors (words) tracked by MPs. The spatio-temporal MP data are used to model individual mobility across regions, and the latent features are used to predict socioeconomic levels under a supervised approach (PMBSEL-sLDA) and an unsupervised approach (PMB-LDA). In predicting regional socioeconomic labels, both approaches exhibit good accuracy ( $R^2 = 0.7802$  for PMBSEL-sLDA and  $R^2 = 0.7188$  for PMB-LDA) and outperform traditional approaches using pre-determined features by about 9% in the best case.

Data recorded by financial systems and smart cards have been used to estimate individual financial status. Singh et al. [498] analyzed a dataset of economic transactions and found an intricate connection between individual financial outcomes and their spatio-temporal traits such as exploration and engagement. Models including such features improve the comparable demographic models by 30% to 49% in predicting future financial difficulties. Lenormand et al. [499] analyzed geotagged credit-card transactions of individuals living in Barcelona and Madrid. They found that younger and older people exhibit differences in traveled distance and purpose of travel, showing the effects of demographic characteristics on mobility patterns. Recently, Zhu et al. [500] inferred economic attributes of urban rail transit passengers from their mobility patterns and personal attributes. They proposed a mobility-to-attribute frame-

work that integrates smart card data (for extracting individual mobility patterns), house prices and shop consumer prices (for estimating economic status). They found that passengers' income is negatively correlated with their commuting distance and transit frequency, suggesting that the daily trip of low-income passengers depends strongly on the metro. In a word, it is a promising method to estimate individual socioeconomic status from financial transactions and transportation modes.

#### 4.2. Employment and performance

Employment and performance are of high relevance to national prosperity, and resignation may result in great losses for companies. Previous studies based on survey data have found that employees' job satisfaction and organizational commitment are predictive to their turnover [501], and job performance is curvilinearly related to turnover [502]. A variety of linear and nonlinear models have been employed to predict unemployment rate based on statistics and jobless claims. Recently, new methods have been proposed to analyze unemployment by utilizing novel data such as Internet search queries, mobile phone (MP) records, and social media (SM) posts. In addition to unemployment, individual and group performances are also of particular interest. As the availability of new data and methods, our understanding of performance has been remarkably improved [503]. In this subsection, we will briefly introduce recent progresses on unemployment prediction and performance analysis.

##### 4.2.1. Search queries indicate unemployment

The Internet provides rich information about people's wants, needs and concerns on a continual basis, and thus it is possible to mine employment-related information and estimate unemployment rate from Internet search queries. Literature have found that some unemployed individuals use the Internet to seek jobs, and human behaviors reflected by search queries are predictive to unemployment data. Ettredge et al. [504] analyzed employment-related Internet search queries recorded by the WordTracker metasearch engines and unemployment data collected from the US Bureau of Labor Statistics. They found that job search queries are positively correlated with the official unemployment data, and the explanatory power of the used regression model decreases with the increase of the lead time. However, the correlation between search queries and unemployment rate holds only for males of age 20 and over. Their work demonstrates the limitation of search queries in predicting unemployment rate.

Based on data from the Google Insights (GI), Askitas and Zimmermann [505] explored the utilization of search queries to predict unemployment rate in Germany. They developed a time-series causality approach to regress monthly unemployment rate against search activity. They found that search queries exhibit strong correlations with unemployment rate, showing that search queries are helpful for economic behavior prediction even under complicated and varying situations. Similarly, Choi and Varian [506] explored how Google search queries of unemployment-related topics, classified by GI and GT [507], are related with the initial jobless claims, a widely accepted indicator of unemployment rate in the US. They found that a GT-based long-term model can help predict initial claims seven days ahead of the official release. D'Amuri [508] leveraged job search queries to predict the quarterly unemployment rate in Italy. The unemployment-related queries are based on GI, and the official unemployment rates are collected from the Italian Labor Force Survey. They found that 1% increase in GI is associated with 0.44% increase in the unemployment rate, and GI-included models perform fairly well.

To comprehensively test GI's predictive power for the US unemployment rate, D'Amuri and Marcucci [509] analyzed over five hundred time-series forecasting models on an out-of-sample forecasting task. They found that models using GI as a leading indicator significantly outperform traditional ones. In particular, GI-augmented models exhibit the mean squared error (MSE) 29% lower when forecasting at one step ahead, and the MSE decreases by 40% when forecasting at three steps ahead. Using unemployment-related Google search queries, Xu et al. [510] proposed a neural-network-based forecasting method for unemployment rate with features mined by several feature selection algorithms. Their method exhibits a higher accuracy than other benchmark methods. Xu et al. [511] further developed a framework to forecast unemployment trend using data mining tools. They extracted employment-related activities from search queries and reduced the data dimension by employing feature selection models. Then, they used neural networks and SVRs [279, 280] to model the relations between search activities and unemployment rate and selected the predictive data mining method with the best feature subset. Their method exhibits an outstanding performance in predicting unemployment rate.

Barreira et al. [512] explored the improvement of the unemployment nowcasting ability based on search queries from four countries in the south-western Europe. They found that the predictive ability of search queries differ

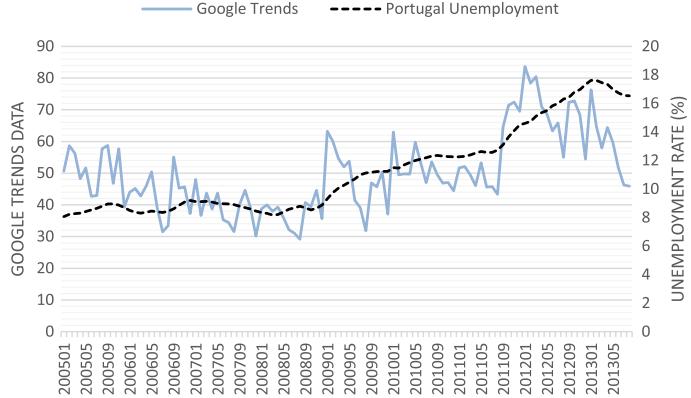


Figure 33: Official unemployment rate versus Google Trends data in Portugal. Figure from [512].

by country and language. Data of GT can improve the nowcasting performance in three out of four considered countries (see Figure 33 for results of Portugal), but the predictive ability is different after considering different out-of-sample periods. In other words, the out-of-sample predictive ability can be improved by GT variables when they have significant in-sample differences. Li et al. [513] proposed an ontology-based Web mining method to better predict unemployment rate based on Google search queries. The domain ontology captures unemployment-related concepts and their semantic relationships and thus contributes to the extraction of useful features. Their method outperforms some baseline methods such as the ARIMA model [514]. Using a time-series approach, Vicente et al. [515] found that Google search queries remain useful in predicting unemployment rate when job destruction is skyrocketing such as the sharp increases in Spanish unemployment due to the economic crisis.

Job-related Google search queries are also predictive to unemployment rate in small and emerging economies. For the Visegrad Group countries, Pavlicek et al. [516] explored the relationship between the intensity of job-related Google search queries and the unemployment rate. Specifically, they used the linear regression model given by

$$\Delta UR_t = \alpha_0 + \alpha_1 \Delta \log(GI)_t + \varepsilon_t, \quad (64)$$

where  $\Delta UR_t$  is the first difference of the unemployment rate,  $\Delta \log(GI)_t$  is the first logarithmic difference of the search queries,  $\alpha_i$  ( $i=0,1$ ) is the regression coefficient, and  $\varepsilon_t$  is the error term at time  $t$ . They found that job-related search queries can track changes of unemployment rate. Further, they proposed a nowcasting model for unemployment rate based on the job-related search queries, as

$$\Delta UR_t = \beta_0 + \sum_{i=1}^L \beta_i \Delta \log(UR)_{t-i} + \sum_{j=0}^L \gamma_j \Delta \log(GI)_{t-j} + \varepsilon_t, \quad (65)$$

where a three-month lag ( $L = 3$ ) in the unemployment rate is assumed available (see Ref. [516] for details). Similar methods have been used to predict unemployment rates in developed and emerging economies such as Italy [517] and Turkey [518]. The best model augmented with Google search queries performs more accurate (38.4% out-of-sample and 47.7% in-sample) than the benchmark autoregressive model in predicting monthly unemployment rates.

Web search queries have also been used to forecast youth unemployment rates. Fondeur and Karamé [519] analyzed unemployment-related Google search queries and official claimant count data in France. They proposed a statistical model that considers the non-stationarity and multiple frequencies in the data and estimated it with the diffuse Kalman filter [520]. They found that search queries can improve unemployment predictions for the 15- to 24-year old French unemployed population. By integrating search queries into the ARIMA model, Kwon and Jung [521] developed a model to predict the Korean youth unemployment rate. Their linear regression model is given by

$$\log(X_t) = \alpha_0 + \alpha_1 \log(X_{t-1}) + \alpha_2 \log(X_{t-12}) + \sum_{k=1}^n \beta_k \log(q_t^k) + \varepsilon_t, \quad (66)$$

where  $\log(X_t)$  is the logarithm of the youth employment rate,  $q_t^k$  is the value of the search volume index,  $\alpha_i$  ( $i=0,1,2$ ) is the regression coefficient,  $\beta_k$  is the coefficient of the query value, and  $\varepsilon_t$  is the error term at time  $t$ . The model exhibits an explanatory power of about 80% for the Korean youth unemployment rate. Naccarato et al. [522] predicted the Italian youth unemployment rate based on Google search queries and official labor data. They employed two time-series models, namely, the ARIMA model using the labor data and the vector autoregressive (VAR) model [523] using both data. They found that VAR outperforms ARIMA in the forecasting error.

Recently, some novel employment-related indicators are proposed for labor markets. Baker and Fradkin [524] developed the Google Job Search Index (GJSI), which is a job search activity index based on Google search queries. GJSI is a useful complement to existing measures as it is available in real time, it has a higher geotemporal resolution, and it suffers less from sampling bias. In addition, GJSI correlates with Google search queries for “jobs”, displays the holiday effects and reveals the search intensities of different groups. Together with the introduction of search queries, large-scale administrative records can also provide new opportunities to understand unemployment. For example, Guerrero and Lopez [525] developed a data-driven model of unemployment dynamics by taking the advantages of fine-grained details of administrative data.

#### 4.2.2. Other sources relevant to unemployment

The availability of large-scale social media (SM) data enables us to better explore unemployment. Based on a large Twitter dataset, Antenucci et al. [526] created a SM signal of job loss, which tracks the initial unemployment insurance claims at medium and high frequencies. They constructed a real-time SM job loss index from the principal components of the collected unemployment-related phrases like “lost my job” in tweets. Results demonstrate the usefulness of SM in constructing indicators of economic activity, more specifically, unemployment status. Proserpio et al. [527] analyzed over 1.2 billion Twitter posts collected from US users during 2010 and 2015. To identify users that gained a new job or lost a job, they searched for tweets containing relevant text strings (e.g., “started my new job” and “I lost my job”). In particular, they explored the relations between psychological well-being and the macroeconomic shock of employment instability. They found that SM is able to capture and track changes in psychological variables over time. They proposed a behavioral model that leverages these changes to predict the US unemployment rate. Their model improves the prediction accuracy by about 25% and 49% compared to the baseline autoregressive model for employed and unemployed samples, respectively.

Based on geotagged tweets collected from Spain, Llorente et al. [528] investigated whether the unemployment incidence information can be revealed from behavioral patterns underlying human mobility, activity and communication. They constructed individual behavioral features by defining four sets of metrics based on tweets (see Figure 34). For SM technology adoption, they found a strong and positive correlation between unemployment and the Twitter penetration rate defined by the fraction of Twitter users in the national census. For the SM activity, the correlation between unemployment and the percentage of tweets is strongly negative in the morning while positive in the afternoon. For the SM content, there is a positive and strong correlation between unemployment and the fraction of misspellers. For the diversity of SM interactions defined by entropy [248], areas with less diverse communication patterns have larger unemployment rates. Further, they employed a simple linear regression model to predict regional unemployment rates using these variables. The model exhibits a strong predictive power  $R^2 = 0.62$  for ages below 25 and  $R^2 = 0.52$  for ages between 25 and 44.

Digital exhaust of human activities left on SM platforms can provide important insights to models of employment-related indicators. Bokányi et al. [529] studied how unemployment and employment statistics of counties in the US are encoded in the daily rhythm of people. They collected 63 million geotagged tweets posted between January and October 2014 from the contiguous US and aggregated them to form a workday tweeting activity pattern with hourly resolution for each county. They found that hourly activities during the daytime (6 am - 8 pm) correlate negatively with unemployment and correlate positively with employment. These results are in accordance with previous findings for Spain that higher morning tweeting activities indicate lower unemployment rates [528]. Further, they decomposed the tweeting pattern of a county into a linear combination of two universal patterns: one group with regular working hours, and the other group who wake up later and stay up until late in the evening. Formally, the predicted activity  $x_i^{(k)}$  of county  $k$  in hour  $i$  is given by

$$x_i^{(k)} = \alpha^{(k)} A_i + (1 - \alpha^{(k)}) B_i, \quad (67)$$

where  $\alpha^{(k)}$  and  $1 - \alpha^{(k)}$  are the mixing proportions for the two universal patterns  $A$  and  $B$ , respectively. Bokanyi et al.

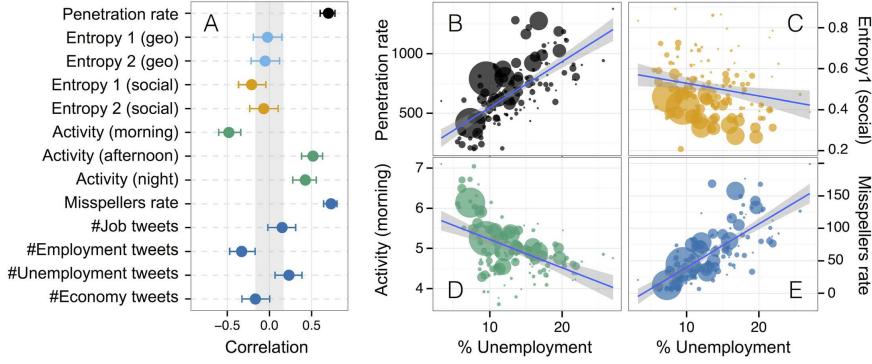


Figure 34: The extracted Twitter metrics and the unemployment rate. (A) Correlation coefficients between behavioral patterns and unemployment rates. Error bars correspond to 95% confidence intervals. (B-E) Relations between the values of 4 selected variables and the unemployment rate in each geographical communities. Size of the points is proportional to the population in each geographical community. Solid lines correspond to linear fits to the data. Figure from [528].

[529] found that the mixing ratio defines a country-specific measure that correlates significantly with unemployment ( $r = -0.34 \pm 0.02$ ) and employment ( $r = 0.46 \pm 0.02$ ). The result demonstrates that daily rhythms of tweets exhibit predictive power for whether individuals have regular working lifestyles.

Turnover of employees may have network effects on the attitudes of stayers. Krackhardt and Porter [530] analyzed a communication network questionnaire and found the snowball effect that turnover occurs in clusters. On the other hand, the structural information of social networks is predictive to individual employment status. Feeley and Barnett [531] studied three social network models of employee turnover and found evidences supporting the Erosion model [531] that employees located on the periphery of a social network are more likely to leave their position. Mossholder et al. [532] analyzed a sample of health care employees and found that network centrality is predictive to turnover. Based on survey data, Feeley [533] found that the employees with large degrees and betweennesses are less likely to leave their jobs. Feeley et al. [534] proposed an improved Erosion model to explain the observed negative correlation between the network centrality and the probability of employee turnover.

Recent availability of large-scale and reliable network data has made it possible to better infer individuals' employment intentions from their centralities in social networks. Based on the social network data collected from a platform used by a Chinese company, Gao et al. [535] built two directed networks: social network (SN) and action network (AN), where links indicate social connections and work-related interactions among employees, respectively. After linking network features to human resource data, they found that the most negatively correlated features with turnover are degree ( $k$ ), out-degree ( $k^{\text{out}}$ ) and k-core value ( $k_s$ ) [536] in AN as well as in-degree ( $k^{\text{in}}$ ), in-strength ( $s^{\text{in}}$ ) and k-core value ( $k_s$ ) in SN (see Figure 35). Moreover, the employees with high in-degrees ( $k^{\text{in}}$ ) are likely to be promoted, and the strongly and positively correlated features with promotion are PageRank index ( $PR$ ) [63] and LeaderRank index ( $LR$ ) [537]. Based on the same data, Yuan et al. [538] further studied the predictability of employee career development. They employed a binary logistic regression model to predict the promotion or resignation of employees based on the network structural features. In the model, the conditional probability of promotion or resignation is given by

$$P(1|\vec{x}) = \frac{1}{1 + e^{-(b_0 + \sum_i^m b_i x_i)}}, \quad (68)$$

where  $\vec{x} = (x_1, \dots, x_m)$  is the vector of structural features, and  $\{b_0, b_1, \dots, b_m\}$  are the coefficients estimated based on the data. After predicting employee career development using the model, they found that features of AN have stronger predictive power for both turnover and promotion than features of SN.

Mobile phone (MP) data have been used to analyze and predict unemployment. Based on call detail records (CDRs) in two European countries, Toole et al. [539] developed a methodology to track employment shocks in nearly real time. They proposed a structural break model to detect mass layoffs based on the drop of calling activity near the plant. The laid-off users are identified by calculating the Bayesian probability weights based on the observed changes

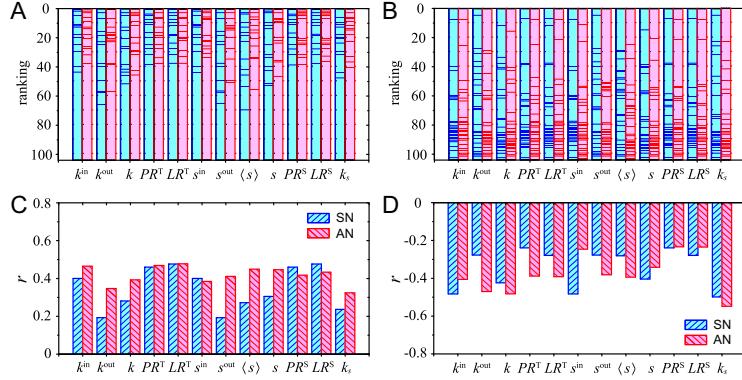


Figure 35: The relations between network centrality measures and employees' career development. (A) and (B) utilize horizontal lines to show the ranking of promoted and resigned employees by centrality measures, respectively. (C) and (D) present the Pearson correlations  $r$  between centrality measures and the probability of promotion and resignation, respectively. The centrality of an individual is measured by in-degree ( $k^{\text{in}}$ ), out-degree ( $k^{\text{out}}$ ), degree ( $k$ ), PageRank index ( $PR^T$ ) and LeaderRank index ( $LR^T$ ) in unweighted networks and the metrics in-strength ( $s^{\text{in}}$ ), out-strength ( $s^{\text{out}}$ ), strength ( $s$ ), average strength ( $\langle s \rangle$ ), PageRank index ( $PR^S$ ) and LeaderRank index ( $LR^S$ ) in weighted networks. The k-core index ( $k_s$ ) [474] is applied in unweighted networks. Figure from [535].

of calling patterns. Formally, user  $i$ 's probability of laid off is given by

$$P(\text{laid off})_i = \frac{\gamma P(\Delta\hat{q}|\Delta q = d)}{\gamma P(\Delta\hat{q}|\Delta q = d) + (1 - \gamma)\gamma P(\Delta\hat{q}|\Delta q = 0)}, \quad (69)$$

where  $\Delta q = q_{\text{pre}} - q_{\text{post}}$  is the difference in the fraction of days on which a user made a call near the plant in 50 days prior to the layoff,  $\gamma$  is the prior that an individual is a non-resident worker at the plant, and  $d$  is the threshold used for the alternative hypothesis. The method correctly identifies the portion of laid-off users. The social interactions of laid-off individuals are less stable and experience significant decline. For example, the total number of calls drops 51%, and the number of outgoing calls drops 54%. The mobility of laid-off individuals generally declines. For example, the number of uniquely visited towers decreases by 17%, and the radius of gyration [490] decreases by 20% relative to the random sample. Further, Toole et al. [539] employed regression models to predict the province-level unemployment rate. Their predictions exhibit high correlations with present unemployment rate ( $r = 0.95$ ) and unemployment rate ( $r = 0.85$ ) one-quarter in the future.

The effectiveness of MP data in predicting individual employment status can be confirmed by external validations. Sundsøy et al. [540] derived a set of features from MP logs in a South-Asian developing country, which reflects users' social, financial and mobility patterns. They employed several machine learning algorithms to predict individual employment status of different profession groups using these features. They found that individual employment status can be predicted with an average accuracy of 0.675 for all profession groups, but the accuracy varies for different groups. The model can also predict whether phone users are unemployed with an accuracy up to 0.735, which is over 30 times better than the random guess. Moreover, they showed that individual employment can be aggregated and mapped geographically to the cell tower level. Almatouq et al. [541] studied whether district-level unemployment rate can be predicted by behavioral features extracted from CDRs in Riyadh. They found that district-level unemployment rate exhibits strong correlation with behavioral features, specifically,  $r = 0.53$  for the number of MP records,  $r = 0.49$  for the percentage of night calls and  $r = -0.40$  for the social diversity [248]. These results indicate that the unemployment rate can be well estimated based on massive MP data in a cost-effective way.

#### 4.2.3. Individual and group performance

Performance of individuals and groups is one of the central concerns of organizations in human resource management. Individual performance is loosely related to the frequently used tokens of success as it is a measure capturing a performer's actions [542]. The quantification of performance remains challenging as it is usually hard to track and analyze individual actions due to the lack of high-quality data. On the other hand, the structure of social and collaboration networks can affect an individual's performance. Through a field study, Sparrowe et al. [543] found

that centrality in advice networks is positively correlated with individual performance, while the density of hindrance network is negatively correlated with group performance. Ahuja et al. [544] examined the determinants of individual performance in virtual R&D groups. They found that network centrality has a stronger predictive power for individual performance than individual characteristics. After analyzing data of surveys representing virtual 35 teams, Kirkman et al. [545] found a positive correlation between team empowerment and performance, while face-to-face interaction can moderate their relationship.

The network effects on performance have been revealed by traditional survey and small-scale data in different contexts. Cross and Cummings [546] built information and awareness networks based on two e-mail surveys. They found that network structures are associated with individual job performance in knowledge-intensive work. In particular, betweenness centrality in both networks is predictive of individual performance. Duch et al. [547] analyzed the performance of soccer players and found that flow centrality is a powerful quantification of individual and team performance. They developed a method of social network analysis to quantify individual performance in the context of soccer. In two studies involving 699 people, Woolley et al. [548] showed that a group's performance can be explained by the collective intelligence factor, which is not determined by individual intelligence but correlated with the proportion of females in the group ( $r = 0.23$ ), the equality in distribution of conversational turn-taking ( $r = -0.41$ ) and the average social sensitivity of group members ( $r = 0.26$ ). Bear and Woolley [549] explored the role of gender diversity in group performance and found that the presence of female can improve group collaboration and practical consequences.

After analyzing social network data collected through questionnaire in China, Cai et al. [550] found that the structure of employees' informal network other than formal network has a significant impact on their performance. In particular, a brokerage's performance is greater affected by direct contacts than indirect contacts. Taking into account the multiplex structure of employee social networks, Cai et al. [551] showed that a nuanced multiplex network model can provide a richer explanation of employee performance than a single-layer model. They built a superimposed multiplex network (SMN) and an unfolded multiplex network (UMN) based on five different categories of employee relationships. They found that different types of social relations have different effects, where employees with high degrees and large eigenvector centralities in the weighted UMN are more likely to perform well. Not only network structure but also team size is predictive to performance. Through an online experiment, Mao et al. [552] found that larger teams have better performance than an equivalent number of independent workers in completing complex tasks, while team members exert lower overall efforts than independent workers.

Recent works have focused on the interactions between group members using novel data collected by digital devices. Pentland [553] analyzed data of many project/industry teams and found that a team's communication pattern is the most important predictor of its performance. They suggested energy, engagement and exploration as three key communication dynamics that affect team performance after analyzing data collected by wearable electronic sensors (see Ref. [554] for details). Moreover, there exists an ideal pattern for each team, and thus a team's performance can be improved by adjusting its communication behavior towards the ideal. Watanabe et al. [555] analyzed data collected by sociometric badges in a call center environment and found that team performance is correlated with the activity level while resting other than the activity level while working. The result suggests a way to improve team performance by enhancing members' face-to-face communications. By analyzing similar data of call centers in China, Tjosvold et al. [556] found that individuals in cooperative teams have higher productivity. The number of phones answered by members of cooperative teams increases about 40% compared to control ones.

Social network data have been used to examine the relations between employees' performance and their social network structure. Gao et al. [535] built two networks, namely, social network (SN) and action network (AN), based on data from a social network platform used by a Chinese company (see also Ref. [538]). They found that centralities of employees in SN, on average, have stronger correlations with performance than those in AN (see Figure 36). The most correlated metrics to employees' performance in SN are in-degree and weighted in-degree with the Pearson correlation  $r \approx 0.48$ . By comparison, the most correlated metrics in AN are in-degree, PageRank and LeaderRank with  $r \approx 0.42$ . De Montjoye et al. [557] explored how network structure affects teams' problem solving abilities in real working environment. They found that only the strongest ties in within-team networks and the members' extended information networks affect the performance of teams.

The analysis and evaluation of researchers' performance is an important part of scientometrics. Recently, an interdisciplinary field named "science of science" (SciSci) [558, 559] emerges, aiming to quantify, understand and predict scientific discoveries and the resulting outcomes of individuals and groups. After analyzing millions of papers and

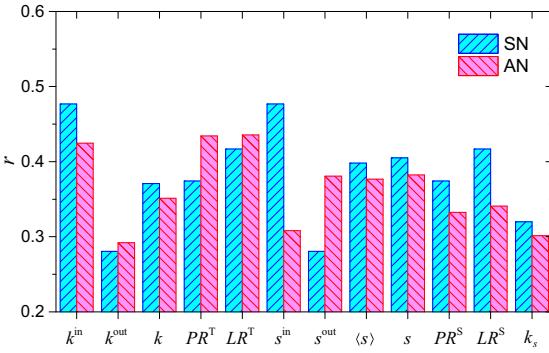


Figure 36: The Pearson correlations between employees' performance and their centralities in the social network (SN) and action network (AN). The centrality of an individual is measured by in-degree ( $k^{\text{in}}$ ), out-degree ( $k^{\text{out}}$ ), degree ( $k$ ), PageRank index ( $PR^T$ ) and LeaderRank index ( $LR^T$ ) in unweighted networks and the metrics in-strength ( $s^{\text{in}}$ ), out-strength ( $s^{\text{out}}$ ), strength ( $s$ ), average strength ( $\langle s \rangle$ ), PageRank index ( $PR^S$ ) and LeaderRank index ( $LR^S$ ) in weighted networks. The k-core index ( $k_s$ ) [474] is applied in unweighted networks. Figure from [535].

patents, Wuchty et al. [560] found the increasing dominance of teams in the production of knowledge. In particular, teams produce more frequently cited and high-impact outcomes than individuals do. Jones et al. [561] further found that cross-university collaborations overall improve paper quality. Collaborations involving top-tier universities produce the highest-quality papers, while weak-weak combinations produce even worse papers than independent researches. De Stefano et al. [562] examined three co-authorship networks of Italian academic statisticians. They found that centrality measures are positively correlated with scientific performance, while local clustering coefficient has a negative influence. Lungeanu et al. [563] analyzed grant proposals and found that successful teams tend to have members with longer tenure, lower institutional tier, more female gender, less prior citation relationships, and so on.

Toward an objective measure of individual scientific performance, many metrics have been considered such as total paper count, citations per paper, impact factors of published journals, and so on. One metric that enjoys a spectacularly quick success is h-index [564], which is defined as the largest integer  $h$  such that there are at least  $h$  papers with  $\geq h$  citations each. Hirsch [565] showed that h-index outperforms other indicators in predicting future scientific achievement of individuals. Later, Radicchi et al. [566] found the universality of citation distributions across disciplines indicated by a universal curve of the relative indicator  $c_f = c/c_0$ , where  $c$  is the citation counts, and  $c_0$  is the average citation per article for discipline  $f$ . As  $c_f$  is an unbiased indicator for citation performance, they introduced a generalized h-index named  $h_f$  index (see Ref. [566] for details) that is suitable for comparing scientists across disciplines. Abbasi et al. [567] proposed the researcher collaboration index (RC-Index) and the community collaboration index (CC-Index) to identify researchers who may be suitable to lead research projects. Till far, there are many variants of h-index to evaluate the performance of scientists (see review article [568]).

Wang et al. [569] developed a mechanistic model to predict a paper's ultimate impact by a single parameter inferred from its early citation history. The unfolded universal temporal pattern in such citation dynamics of papers can be used to evaluate scientific impacts of individuals. Sinatra et al. [570] quantified the evolution of individual scientific impact based on publication records and career profiles. They developed a quantitative model of scientific impact and proposed a factor  $Q$  to capture a scientist's sustained ability to publish high-impact papers. Interestingly, the factor  $Q$  is a fingerprint for scientists and independent to their career stages. The  $Q$ -model can predict future time evolution of individual scientific impact. After analyzing scientists' career trajectories, Deville et al. [571] found that individuals moved from elite to lower-rank institutions tend to experience modest decrease in scientific performance, while movements towards elite institutions do not bring subsequent performance gain. Shen and Barabási [572] proposed a credit allocation algorithm that can accurately measure the relative credits for different coauthors. Jia et al. [573] analyzed publication records and found an exponential distribution of changes in research interests. They further developed a random-walk-based model to accurately reproduce the empirical observations.

Researchers have also tried to connect students' educational achievements with their behavioral patterns. Based on passive sensing data from smart phones, Wang et al. [574] explored individual behavioral differences among a group of students with different performance. They found a number of important behavioral factors that are significantly correlated with term and cumulative GPA, such as conversational interaction, class attendance, and studying hours.

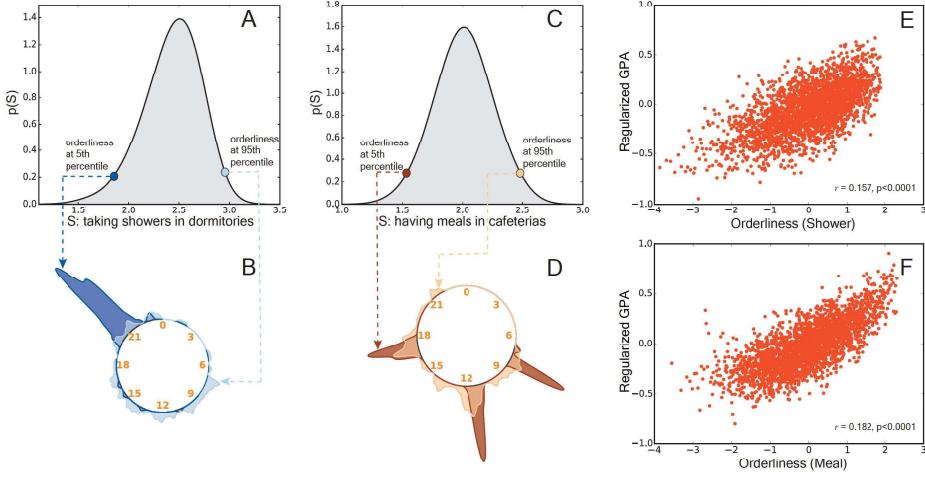


Figure 37: Distributions of actual entropies of students in (A) taking showers and (C) having meals. The behavioral clocks of two students at the 5th percentile and the 95th percentile are shown for (B) taking showers and (D) having meals. Correlations between regularized GPA and (E) regularized orderliness (Shower) as well as (F) regularized orderliness (Meal). The corresponding Spearman's rank correlation coefficients  $r$  and the level of statistical significance  $p$  are also shown in the plots. Figure after [575].

Using these behavioral features, a linear regression model can predict cumulative GPA with  $r = 0.81$ . Using behavioral records collected by students' smart cards, Cao et al. [575, 576] quantified the relations between behavioral patterns and academic performance. They introduce orderliness, a novel metric based on the actual entropy [577, 578], to measure the regularity of each student's campus lifestyle based on the temporal records of having meals and taking showers (see Figure 37). They found that orderliness exhibits a significantly positive correlation with GPA, and it can remarkably improve the prediction accuracy of students' academic performance at the presence of other behavioral indicators. Similarly, Yao et al. [579] collected longitudinal behavioral data of 6,597 students through smart cards and proposed three behavioral features, namely, orderliness, diligence, and sleep patterns. They further built a multi-task predictive framework that can well predict student's academic performance by utilizing proposed behavioral features.

#### 4.3. Demographics and personal variables

Demographic attributes of individuals have remarkable effects on their socioeconomic status, while traditional methods of individual profiling based on surveys and censuses are costly and follow a long-time delay. Recently, data from novel sources such as social media (SM) and mobile phones (MPs) have been used alternatively to predict individual demographic attributes. Moreover, individual behaviors on social networking platforms have been used to estimate individual mental states such as emotion, depression and suicidal intent. Meanwhile, these novel data sources have also been leveraged to predict personality and evaluate reputation of individuals.

##### 4.3.1. Demographic inference

Understanding demographics of individuals has important applications in estimating socioeconomic outcomes. Beyond traditionally used census, data from MPs and Twitter have been used to infer demographic attributes such as gender and age. Using a rich set of features extracted from tweets, Rao et al. [580] developed a stacked-SVM-based classification algorithm to classify latent user attributes, giving an accuracy 0.723 on gender classification. The algorithm outperforms the baseline ngram-only model with accuracy 0.687 and the SVM-based binary classifier with accuracy 0.718 [581]. Based on a Twitter dataset labeled with gender, Burger et al. [582] applied some statistical models to predict gender using features of both word- and character-level ngrams. They found that the most informative feature is a user's full name, which provides an accuracy 0.891 in gender classification. Moreover, tweets contribute more than user description in predicting a user's gender. Test classifier with all features extracted from user profile and texts exhibits an accuracy about 0.92.

By calculating gender-name association scores, Liu and Ruths [583] explored the link between gender and first name in English tweets. They found that including first name can improve the accuracy of gender inference by about

20%. They further developed a method to identify gender-labels without analyzing user profile or textual content. Ciot et al. [584] assessed gender inference methods based on non-English tweets. They found that existing machinery can address the gender inference problem, and including language-specific features can make accuracy gains. Volkova et al. [585] applied machine learning and natural language processing techniques to predict personal attributes. They trained log-linear models using lexical features extracted from 200 tweets per user profile and identified male gender with an accuracy 0.8. Culotta et al. [586] predicted the demographics of Twitter users based on whom they follow. Specifically, they labeled demographics of visitors to over 1,500 websites. Then, they predicted demographics of Twitter users by a regression model using the information of users following the websites' accounts on Twitter. Montasser and Kifer [587] developed a method to predict a region's demographics based on the characteristics of geotagged tweets in that region. Using lexical features extracted from tweets, their method predicts census-based race data at the block level with an average accuracy 0.692.

Content and network features of SM users are predictive of their occupations. Huang et al. [588] developed an integration framework to infer users' occupations from their social activities on Weibo. They proposed a content model to identify beneficial content features and used a network model to identify user latent affiliations. Their model that integrates both network and content exhibits an accuracy 0.80 on occupation inference. Preotiuc-Pietro et al. [589] employed linear and nonlinear models to predict a user's occupational class based on latent features extracted from tweets. They found that text features can improve the performance, and the best model can give an accuracy of over 0.50 for a 9-way occupation classification. Sloan et al. [590] developed two methods to extract social class of occupation from the profiles of UK Twitter users. Their methods can identify certain occupational groups such as professionals.

MP data have been used to infer demographic information. Frias-Martinez et al. [591] analyzed call detail records (CDRs) and found that male and female users are significantly different in behavioral and social variables such as duration of calls and degree in social networks. They proposed a semi-supervised classification algorithm that can identify gender with an accuracy up to 0.80. From CDRs, Herrera-Yagüe and Zufiria [592] extracted 22 features that are most relevant to gender. They found that females tend to have a higher average call length, a larger median of call duration and more messages sent per relationship. They tested several machine learning schemes and found that SVM using call features performs the best with an accuracy over 0.60 for gender prediction. After analyzing daily communication patterns extracted from one billion CDRs, Dong et al. [593] found that female users pay more attention to cross-generation interactions. They proposed a factor graph model name WhoAmI that accounts for the interrelations to infer gender and age. The WhoAmI outperforms benchmark methods by about 10% in terms of F1 measure. Dong et al. [594] generalized WhoAmI to infer any number of interrelated attributes and proposed a coupled WhoAmI method to predict demographics across two mobile operators. Their new methods exhibit accuracies up to 0.80 and 0.73 in predicting gender and age, respectively.

After extracting user behavioral and social variables from Mexican CDRs, Sarraute et al. [595] evaluated several machine learning algorithms for gender prediction. They found that the logistic regression and linear SVM perform the best with an accuracy 0.814. Including individual calling patterns and communication network structures, machine learning algorithms can also predict other demographic variables such as age. Jahani et al. [596] developed a framework to predict individual characteristics based on over 1400 features extracted from CDRs. They found that machine learning algorithms trained with only 10,000 users are sufficient to predict gender with an accuracy up to 0.884 in a south Asian developing country and with an accuracy up to 0.797 in an European developed country. The performance can be slightly improved by increasing the minimum required days that users are active per week (see Figure 38). Moreover, their method can be used to predict other demographic variables such as age with  $R^2 = 0.47$  in a multi-classification task.

Based on GPS data collected from MPs, Akter and Holder [597] proposed a graphical-feature-based framework to improve demographic prediction. They constructed user-wise networks, in which nodes are location categories and edges represent users' movements between location categories. Then, they extracted relevant graphical features from the graph representation and trained a SVM to identify gender. Their method with the optimal set of features exhibits an accuracy 0.8599 in gender classification. Wang et al. [598] analyzed data of recorded MP connections to access points (APs) in two campuses and found that users' demographic attributes can be solely inferred from their spatiotemporal AP-trajectories. They developed a method named Sinfer to infer the social network of users based on the co-occurrence events of AP-trajectories. Further, they proposed a tensor-factorization-based method named Dinfer to predict users' demographic attributes using the social network learned by Sinfer. Their method gives an F1

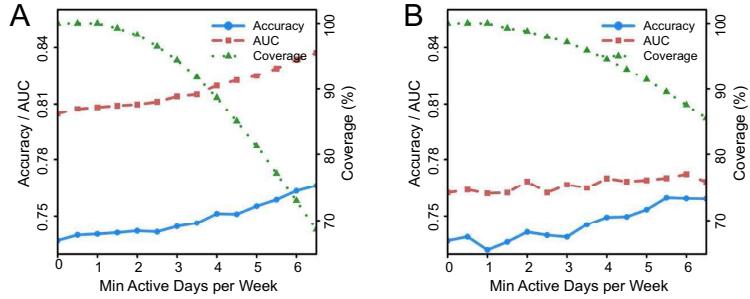


Figure 38: Accuracy and AUC of gender prediction in (A) the European and (B) the south Asian country as a function of the minimum required active days per week. As the minimum required threshold increases, the coverage of the CDR data decreases while the algorithm performance measured by the accuracy and AUC improves slightly. Figure from [596].

measure about 0.7 in gender prediction. Felbo et al. [599] developed the convolutional network (ConvNet) architecture to transform MP data into high-level features for each week and then aggregated patterns across weeks by reusing the same convolutional filters. They designed a 2-step model (ConvNet-SVM) using an SVM with a radial basis function kernel, which slightly outperforms the state-of-the-art method, with accuracies 0.797 and 0.631 in predicting gender and age, respectively.

Digital traces on other platforms have also been used to infer demographic information. Kosinski et al. [139] analyzed a dataset of Facebook Likes and found that users' online behaviors are predictive of their personal attributes such as ethnicity, age and gender. Zhong et al. [600] analyzed profiles of 159,530 SM users and found that a variety of demographic attributes can be inferred from check-ins, including education background, marital status, gender, and age. They proposed a location-to-profile framework that outperforms benchmark methods in inferring user profile. Using data of Facebook and Pokec social networks, Lin et al. [601] proposed an algorithm that can predict user profile with a high accuracy. Recently, Ren et al. [602] studied demographic prediction using different types of data including cyber, physical and social behaviors. They found that including cyber-physical-social behaviors into an SVM model can significantly increase the accuracy for gender prediction. Readers are encouraged to read a recent review about demographic attribute prediction based on online digital traces [603].

#### 4.3.2. Personality analysis

Personality is usually quantified by the five-factor model (FFM) [604, 605], which suggests five broad dimensions to characterize human personality: conscientiousness, agreeableness, neuroticism, openness, and extraversion. Yet, FFM-based analyses are mainly driven by personality survey data. Ross et al. [606] explored the association between personality and Facebook usage patterns. They found that neuroticisms prefer to post photos on their profiles, and extravertions tend to report membership in more groups. Correa et al. [607] investigated the relations between social media (SM) activity and three personality factors (extraversion, neuroticism, and openness). They found that extraversion and openness are positively correlated with SM activity, while neuroticism is negatively correlated with SM activity. Yet, these results are affected by age and gender. For example, extraversion is strongly correlated with SM activity among the young adult cohort.

Golbeck et al. [608] demonstrated that personality of users can be inferred from their Facebook profiles. They found a positive correlation ( $r = 0.264$ ) between conscientiousness and words surrounding social processes. They employed two machine learning algorithms to predict personality factors using 74 features and found that each factor can be predicted within on average 11% of its actual value. Bachrach et al. [609] examined the relations between personality of users and their Facebook profiles based on data of 180,000 users. They found that openness and neuroticism are positively correlated with the number of likes and group associations of users, while the correlation is negative for conscientiousness. They developed a multivariate linear regression to predict each factor based on multiple profile features. Their model exhibits reasonable prediction accuracy for some factors such as extraversion ( $R^2 = 0.33$ ) and neuroticism ( $R^2 = 0.26$ ). They further applied several more sophisticated machine learning methods but found that the improvement of predication accuracy is very limited.

After surveying the personality and the Facebook behaviors of 184 undergraduates, Seidman [610] found that

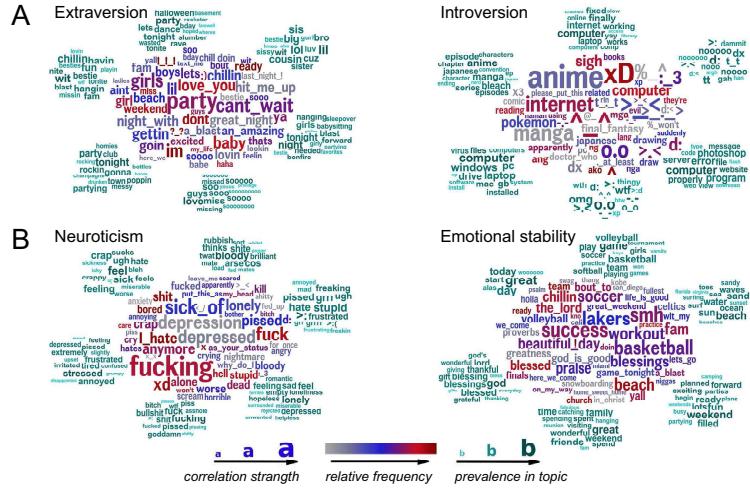


Figure 39: Word clouds for extraversion and neuroticism. (A) Language distinguishing extraversion (left, e.g., “party”) from introversion (right, e.g., “computer”). (B) Language distinguishing neuroticism (left, e.g. “hate”) from emotional stability (right, e.g., “blessed”). Figure after [611].

belongingness-related behaviors can be best predicted by high agreeableness and neuroticism, and more frequent use of Facebook is highly correlated with extraversion. Self-presentational behaviors and motivations can be predicted by low conscientiousness and high neuroticism, while the motivation to express self-aspects can mediate the relationship between neuroticism and self-disclosure. Schwartz et al. [611] analyzed millions of Facebook messages and found that language features can distinguish demographic and psychological attributes (see Figure 39). They extracted 700 million instances by developing a method of open-vocabulary analysis and then linked them with personality. They found that neurotic people disproportionately use the phrase “sick of”, while extraverts tend to use social words such as “party”. Further, they employed the ridge regression [612] to predict each factor using the open-vocabulary features. Their method exhibits fine performance, for example, the accuracy of predicting openness is  $R^2 \approx 0.18$ .

Data from Twitter have also been used to study individual personality. Preotiuc-Pietro et al. [613] developed a multi-modal approach to quantify observable user behaviors on Twitter using natural language processing and image analysis. After extracting online behavioral features capturing text use, profile picture posting and general profile information, they explored how these features are related to dark traits of human personality [614] including narcissism, machiavellianism and psychopathy. They found that trust and positive sentiments are correlated with narcissism, a wide range of negativity and morbidity are correlated with psychopathy, and allowing advertisements in personal communications is correlated with machiavellianism. Further, they built a linear regression model to predict the dark triad traits with a robust performance of Pearson coefficient  $r \approx 0.25$  for all three traits. Similarly, semantic content of Facebook updates has also been used to predict the dark traits [615]. Guntuku et al. [616] explored the relations between personality and online image posting based on about 1.5 million images posted on Twitter. They found that image content of posts and likes can predict the big five personality traits. For example, the performance measured by the Pearson correlations are 0.530 and 0.566 in predicting neuroticism and conscientiousness, respectively.

Based on tweets and profile images of over 66,000 Twitter users, Liu et al. [617] estimated the big five personality. They used the state-of-the-art text prediction methods developed by Schwartz et al. [611] to predict personality based on tweets and employed deep learning methods (see Ref. [618]) to extract facial features from profile pictures. After linking the text-based personality to the facial features, they found that users using appealing profile images are likely to have high openness, profile images with faces are good indicators of higher conscientiousness, and neurotic people have a strong tendency not to present faces. Moreover, these interpretable visual features can be harnessed to predict personality traits with robust accuracy, for example, the Pearson correlation is 0.189 in predicting conscientiousness. Segalin et al. [619] analyzed the effectiveness of visual features on the prediction of personality traits based on Facebook profile pictures. They found that agreeable individuals and extroverts tend to use warm colored pictures, while neurotic individuals tend to post pictures of indoor places. Based on these visual features, they developed a classification method to predict personality traits, exhibiting a mean classification accuracy of about 0.60 for each of

the five traits.

#### 4.3.3. Online reputation evaluation

Reputation has received considerable attention recently in a variety of disciplines [620]. Mui et al. [621] provided an overview of reputation studies across disciplines by summarizing existing notions of reputation and discussing their advantages. Among existing notions, individual reputation, referring to the judgment of an individual's impression by others, is a valuable asset in online social lives. For example, Zacharia et al. [622] studies reputation in an on-line community and found that reputation is related to ratings received by an individual from others. So far, there have been many computational reputation models [623] and many reputation systems for online services [624, 625]. In the following, we will introduce some basic notations for online rating systems and review recent literature that evaluate user reputation.

An online rating system consisting of  $m$  users and  $n$  objects can be described by a weighed bipartite network  $G = \{U, O, E\}$ , where  $U = \{U_1, U_2, \dots, U_m\}$ ,  $O = \{O_1, O_2, \dots, O_n\}$  and  $E = \{E_1, E_2, \dots, E_l\}$  are sets of users, objects and ratings, respectively. Naturally, the bipartite network can be represented by a rating matrix  $A$ , whose element  $A_{i\alpha}$  is the rating given by user  $i$  to object  $\alpha$ . Here, Greek and Latin letters are used for object-related and user-related indices, respectively. Reputation evaluation model aims to assign user  $i$  with reputation  $R_i$  by analyzing the bipartite network  $G$ . Ranking-based reputation evaluation methods can be roughly classified into two categories, namely, quality-based ranking methods and group-based ranking methods.

The quality-based ranking methods assume that a most objective rating can best reflect true quality  $Q_\alpha$  of an object  $\alpha$ . Due to the lack of true quality information, the weighted average rating is used as a proxy. Formally, the estimated object quality  $\hat{Q}_\alpha$  of object  $\alpha$  is given by

$$\hat{Q}_\alpha = \frac{\sum_{i \in U_\alpha} R_i A_{i\alpha}}{\sum_{i \in U_\alpha} R_i}, \quad (70)$$

where  $U_\alpha$  is the set of users who have rated object  $\alpha$ , and  $R_i$  is the reputation of user  $i$ . The most straightforward method is the iterative refinement (IR) [626], which calculates user reputation and object quality in an iterative way. The reputation of a user is inversely proportional to the difference between the vectors of the user's ratings and the estimated object qualities, as

$$IR_i = \left( \frac{1}{k_i} \sum_{\alpha \in O_i} (A_{i\alpha} - \hat{Q}_\alpha)^2 + \varepsilon \right)^{-\beta}, \quad (71)$$

where  $O_i$  is the set of objects rated by user  $i$ ,  $k_i$  is the degree of user  $i$ ,  $\beta$  is a tunable parameter, and  $\varepsilon$  is a small quantify to avoid zero value of the summation. User reputation is initialized as  $IR_i = 1/n$  and iteratively updated by Eq. (70) and Eq. (71) (setting  $R_i \leftarrow IR_i$  in Eq. (70)) until user reputation  $IR_i$  and object quality  $\hat{Q}_\alpha$  converge. Note that,  $IR_i$  should be normalized after every iteration.

Zhou et al. [94] proposed a correlation-based ranking (CR) method, which is more robust under spamming attacks. In the CR method, reputation of a user is iteratively determined by the correlation between the vectors of the ratings  $A$  and the estimated object qualities  $\hat{Q}$ . For user  $i$ , a so-called temporal reputation  $TR_i$  is calculated by the Pearson correlation, as

$$TR_i = \frac{1}{k_i} \sum_{\alpha \in O_i} \left( \frac{A_{i\alpha} - \mu(A_i)}{\sigma(A_i)} \right) \left( \frac{\hat{Q}_\alpha - \mu(\hat{Q}_i)}{\sigma(\hat{Q}_i)} \right), \quad (72)$$

where  $\mu(A'_i) = \sum_\alpha A'_{i\alpha}/k_i$  and  $\sigma(A'_i) = \sqrt{\sum_\alpha (A'_{i\alpha} - \mu(A'_i))^2/k_i}$  are the mean value and standard deviation of  $A'_i$ , respectively. The reputation is set as  $CR_i = 0$  if  $TR_i$  is smaller than 0, and  $CR_i = TR_i$  otherwise. In this process, user reputation  $CR_i$  and object quality  $\hat{Q}_\alpha$  are iteratively updated according to Eq. (70) and Eq. (72) (setting  $R_i \leftarrow CR_i$  in Eq. (70)). To start the iteration, user reputation is initialized as  $CR_i = k_i/n$ , where  $k_i$  is the degree of user  $i$ .

Under the CR framework, Liao et al. [627] proposed an iterative algorithm with reputation redistribution (IARR) by enhancing the influences of users with high reputation. The user reputation in IARR is calculated by nonlinearly redistributing the reputation given by CR, as

$$IARR_i = CR_i^\theta \cdot \frac{\sum_j CR_j}{\sum_j CR_j^\theta}, \quad (73)$$

where  $\theta$  is a tunable parameter. Note that, IARR degenerates to CR when  $\theta = 1$ . Liao et al. [627] further proposed an enhanced iterative algorithm named IARR2 by introducing two penalty factors. Specifically, they modified the calculation of object quality  $\hat{Q}$  in Eq. (70) by

$$\hat{Q}'_\alpha = \max_{i \in U_\alpha} \{R_i\} \cdot \hat{Q}_\alpha, \quad (74)$$

and they modified the calculation of temporal reputation  $TR$  in Eq. (72) by

$$TR'_i = \frac{\log k_i}{\max_j \{\log k_j\}} \cdot TR_i. \quad (75)$$

In a word, IARR eliminates the noisy information by reducing the influence of users with low reputation, while IARR2 gradually filters out the influence of less reliable users in the iterations.

Liu et al. [628] proposed an improved iterative algorithm (IRUA) to rank user reputation by taking into account the role of users' activity patterns. Specifically, IRUA considers the maximum degree of the users who have rated an object and estimates the quality of object  $\alpha$  by

$$\hat{Q}''_\alpha = \max_{i \in U_\alpha} \left\{ \frac{k_i}{n} \right\} \cdot \hat{Q}_\alpha, \quad (76)$$

where  $k_i$  is the degree of user  $i$ , and  $n$  is the total number of objects. The reputation of user  $i$  is updated by considering both the temporal reputation  $TR_i$  calculated by Eq. (72) and the user degree  $k_i$ . The reputation of user  $i$  is given by

$$IRUA_i = \begin{cases} \left( \frac{k_i}{k_{\max}} \right)^\theta \cdot TR_i, & TR_i \geq 0 \\ 0, & TR_i < 0 \end{cases} \quad (77)$$

Here,  $k_{\max}$  is the maximum degree of all users, and  $\theta$  is a tunable parameter that enhances the reputation of large-degree users when  $\theta > 0$ .

Recently, Liu et al. [629] proposed a parameter-free reputation ranking method based on the beta probability distribution (RBPD). Firstly, a rating  $A_{i\alpha}$  is transformed to the extent of fanciness  $A'_{i\alpha}$  by a normalization method:  $A'_{i\alpha} = 2(A_{i\alpha} - A_i^{\min})/(A_i^{\max} - A_i^{\min})$  if  $A_i^{\max} \neq A_i^{\min}$ , and  $A'_{i\alpha} = 0$  otherwise, where  $A_i^{\max}$  and  $A_i^{\min}$  are respectively the maximum and minimum ratings given by user  $i$ . Then, the Bayesian analysis is used to determine the number of fair ratings  $s$  and unfair ratings  $f$  given by user  $i$  based on  $A'_{i\alpha}$ . A rating is fair if it is consistent with the majority of other users' opinions on the corresponding object (see Ref. [629] for details). Finally, the RBPD method defines the user reputation as the probability of giving fair ratings to objects. Formally, the reputation of user  $i$  is given by

$$RBPD_i = \frac{s_i + 1}{k_i + 2}, \quad (78)$$

where  $s_i$  is the number of fair ratings given by user  $i$ . The RBPD method updates  $RBPD_i$  by Eq. (78) and  $\hat{Q}_\alpha$  by Eq. (74) in an iterative manner until object qualities become stable.

The group-based ranking methods define user reputation by the group sizes after grouping all users by their rating similarities. Instead of following the traditional assumption that each object has only one rating that best reflects its quality, the group-based ranking methods underlie that one object should accept multiple reasonable ratings since the users' ratings are subjective and can be affected by many factors [630]. For a discrete rating system with  $\Omega = \{\omega_1, \omega_2, \dots, \omega_z\}$ , Gao et al. [631] proposed a group-based ranking (GR) method that involves the following steps: (i) group users according to their ratings; (ii) calculate the group size matrix  $\Lambda$ , where  $\Lambda_{sa}$  is the number of users who rated object  $\alpha$  with rating  $\omega_s$ ; (iii) build the rating-rewarding matrix  $\Lambda_{sa}^* = \Lambda_{sa}/k_\alpha$ , where  $k_\alpha$  is the degree of object  $\alpha$ ; (iv) map the original rating matrix  $A$  to the rewarding matrix  $A'$ , where  $A'_{i\alpha} = \Lambda_{sa}^*$  if  $A_{i\alpha} = \omega_s$ . If  $A_{i\alpha} = 0$  (i.e., the user  $i$  didn't rate the object  $\alpha$ ), the value of  $A'_{i\alpha}$  is null and should be ignored in the following calculation; (v) calculate the reputation  $GR_i$  of user  $i$  by dividing the mean value of  $A'_i$  by its standard deviation. Formally,  $GR_i$  is defined as

$$GR_i = \frac{\mu(A'_i)}{\sigma(A'_i)}, \quad (79)$$

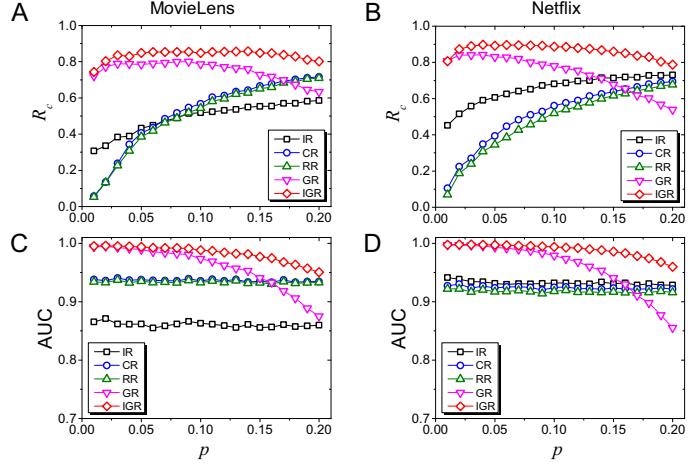


Figure 40: Performance of different reputation ranking methods. All methods are evaluated on the MovieLens (A and C) and Netflix (B and D) datasets, in which  $p$  ratio of ground truthing low-reputation users are assigned with the minimum and maximum allowable ratings. Two evaluation metrics are used, namely, recall ( $R_c$ ) and AUC. Figure from [632].

where  $\mu(\cdot)$  and  $\sigma(\cdot)$  are mean value and standard deviation, respectively. As presented, GR assigns users with high reputation if they always fall into large rating groups.

Later, Gao and Zhou [632] proposed an iterative group-based ranking (IGR) method by introducing an iterative reputation-allocation into the GR method. In IGR, the sizes of user rating groups are weighted by the reputation of users, where ratings from users with higher reputation have larger influences. Formally, the weighted sizes of user rating groups are calculated by

$$\Lambda_{sa} = \sum_{i=1}^m IGR_i \cdot B_{sa}^{(i)}, \quad (80)$$

where  $IGR_i$  is the reputation of user  $i$  calculated in the previous step,  $B_{sa}^{(i)}$  is the rating-object matrix of user  $i$ , and  $m$  is the total number of users. Here, the rating-object matrix  $B_{sa}^{(i)} = 1$  if  $A_{ia} = \omega_s$ , and  $B_{sa}^{(i)} = 0$  otherwise (see Ref. [632] for details). Following GR, the rating-rewarding matrix is calculated by  $\Lambda_{sa}^* = \Lambda_{sa}/k_a$ , and the rewarding matrix is mapped by  $A'_{ia} = \Lambda_{sa}^*$ . The user reputation  $IGR_i$  is re-allocated via  $IGR_i = \mu(A'_i)/\sigma(A'_i)$ . IGR iteratively updates the weighted group sizes by Eq. (80) and the user reputation by Eq. (79) until convergence. The reputation of every user  $i$  is initialized with the same value ( $IGR_i = 1$ ). Experimental results on the MovieLens and Netflix datasets demonstrate the improved accuracy and robustness of IGR in ranking low reputation users (see Figure 40).

Recently, Dai et al. [633] proposed a group-based ranking method based on the user preference, named PGR method, which is a variant of the original GR method. PGR is based on the idea that the preferences of online users are diverse when they give ratings to objects. For example, large-degree users tend to give low ratings [634]. Different from GR where users are grouped by their ratings, PGR divides users into groups by their rating preferences. First, a user rating  $A_{ia}$  is transformed to a mapped rating  $A'_{ia}$  by a normalization method:  $A'_{ia} = (A_{ia} - \mu(A_i))/(A_i^{\max} - A_i^{\min})$  if  $A_i^{\max} \neq A_i^{\min}$ , and  $A'_{ia} = 0$  otherwise, where  $\mu(A_i)$  is the average rating given by user  $i$ . The matrix  $A'$  measures user preference of giving high or low ratings. Then, a new rating matrix  $A''$  is constructed by transforming  $A'_{ia}$  to new ratings  $A''_{ia}$ , where a linear mapping is used to ensure that the sets of rating values for  $A''$  and  $A$  are the same (see Ref. [633] for details). Finally, the user reputation  $PGR$  is calculated under the framework of GR (replacing  $A$  by  $A''$ ).

Besides these aforementioned quality- and group-based methods, scientists have proposed some other user reputation evaluation methods for online rating systems. For example, Fouss et al. [635] proposed a probabilistic reputation model based on transaction ratings. Liao et al. [636] developed a general ranking method to evaluate user reputation in online communities. Li et al. [637] proposed a topic-biased user reputation model for online rating systems. Li et al. [638] reviewed some reputation ranking methods and further proposed six new reputation-based algorithms which exhibit better effectiveness, efficiency and robustness. Indeed, many factors can affect the performance of reputation ranking methods such as the resolution of ratings. After analyzing the effects of discrete and continuous ratings, Medo

and Wakeling [639] found that the overall performance of reputation ranking can be improved by increasing noise in ratings when the rating resolution is low. Recently, Liao et al. [640] reviewed both static and time-aware ranking algorithms and emphasized the benefits by including the temporal dimension.

#### 4.3.4. Emotion and health analysis

People are accustomed to express feelings through social media (SM). Texts posted on SM have been used to track the emotion intensity of individuals. Thelwall et al. [641] examined the extent to which emotion is present in online comments. They classified positive and negative emotions of an initial set of 2,600 human-classified comments from US users in MySpace. They found that two thirds of the comments express positive emotions and a minority contain negative emotions, showing that MySpace is an emotion-rich environment. Thelwall et al. [642] proposed a new method named SentiStrength, which applies machine learning approaches to extract positive and negative sentiment strength from informal text. Applied to MySpace comments, SentiStrength can predict positive and negative emotions with accuracies 0.606 and 0.728, respectively. The result demonstrates the feasibility of using SM to predict emotions of online users.

Twitter provides a rich data source for individual emotional inference. Pak and Paroubek [643] performed a linguistic analysis of tweets and trained a sentiment classifier to determine positive, neutral and negative sentiments for a document. They found that SM users describe different emotions using different syntactic structures. Their methods using N-gram and POS-tags (part-of-speech) as features are efficient in identifying emotional sentiments. Mislove et al. [644] developed color-coded cartograms to track the mood of each state in the US based on over 300 million tweets. They found that the highest level of happiness occurs in the early morning and late evening, weekends are happier than weekdays, and the west coast is happier than the east coast. Bollen et al. [645] extracted six mood states (tension, depression, anger, vigor, fatigue, and confusion) from tweets. They found that popular events can affect many dimensions of public moods. These results suggest that SM can be used to track real-time emotive landscape and trend.

Wang et al. [646] created an emotion-labeled dataset of about 2.5 million tweets and identified emotions by training two machine learning algorithms, namely, LibLinear [647] and multinomial naive Bayes [648]. They found that the most effective features are unigrams, big rams, sentiment/emotion-bearing words, and POS information. With a training data containing about 2 million tweets, the algorithms can achieve the highest accuracy 0.6557 on emotion identification. Larsen et al. [649] constructed a so-called “We Feel” system to analyze variations in emotional expression on Twitter. The system collected 2.73 billion emotional tweets over a 12-week period and classified emotional words into six primary emotion categories with 25 subgroups of secondary emotions. The system can detect emotional responses to significant events and identify depression burdens from emotions expressed on Twitter. Jones et al. [650] found an increase in post-event negative emotion expression on Twitter after mass violence, suggesting the effects of traumatic events on user emotion.

Recently, Mohammad and Bravomarquez [651] annotated a dataset of tweets with emotion intensities, where the best-worst scaling technique is used to improve the annotation consistency [652]. They found that emotion-word hashtags often impact emotion intensity. Further, they developed a regression model to explore the usefulness of features in emotion intensity detection and found that word embedding and lexicon features are the best indicators ( $r = 0.66$ ). Madisetty and Desarkar [653] employed an ensemble of three machine learning methods to determine the emotion intensity of tweets. Specifically, SVR [279, 280] uses lexicon and word embedding features, CNN [654] uses word embedding features, and XGBoost [655] uses word n-gram and char n-gram features. The ensemble method outperforms some baseline methods by giving the Spearman rank correlation 0.725.

Data from other online SM platforms, such as Facebook and Weibo, can also be used to analyze emotion. Settanni and Marengo [656] collected self-report measures of stress, anxiety and depression of over 200 adult Facebook users from North Italy, and explored the relationship between users’ posts on Facebook and their emotional well-being. Through correlation analyses, they found that individuals who have higher levels of anxiety and depression will express negative emotions on Facebook more frequently. Moreover, the use of positive emotions has a negative correlation with the level of stress. By sentimentally analyzing 210 million geotagged tweets collected from Weibo, Zheng et al. [657] constructed a daily city-level happiness index and further explored its relation to daily local PM2.5 concentrations. After exporting the data for 144 Chinese cities in 2014, they found that the happiness index is negatively correlated with PM2.5 concentrations (see Figure 41). They work demonstrates the possibility of capturing users’ emotional expressions and providing real-time feedback about life concerns using SM data.

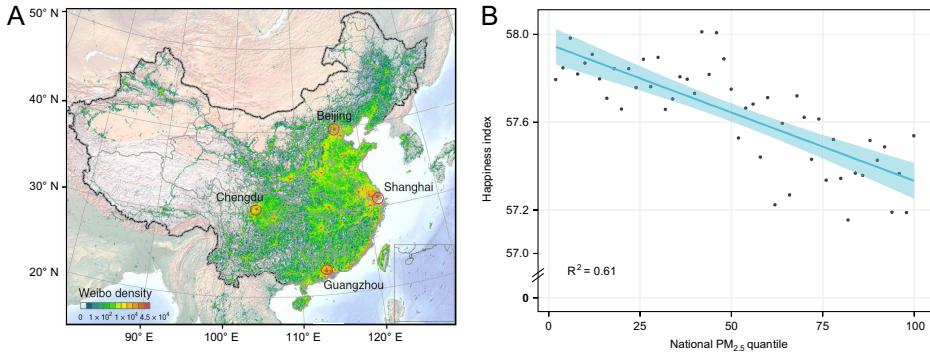


Figure 41: The geography of Weibo tweets, and the relationship between PM2.5 concentration and happiness index. (A) The spatial distribution of geotagged Weibo tweets in China. (B) The relationship between PM2.5 concentration and the happiness index. The happiness index ranges from 0 to 100. The PM2.5 concentration is divided into 50 groups, and the median happiness index value for each group is represented as a dot. The dots are fitted by the downwards sloping line with the blue shaded area represents the 95% confidence interval. Figure from [657].

Rich information provided by SM has been used to study mental health. Coppersmith et al. [658] explored the linguistic signals relevant to specific disorders and mental health. They gathered mental-illnesses-related data and replicated previous findings based on the linguistic inquiry word count (LIWC) [659], suggesting the relevance of tweets to mental health. Coppersmith et al. [660] further built machine learning classifiers using self-reported statements of diagnosis to identify Twitter users with ten serious mental conditions such as post-traumatic stress disorder (PTSD), generalized anxiety disorder (Anxiety), and eating disorders (Eating). Their classifiers exhibit reasonable performance for most mental conditions, for example, the highest precision is 0.85 for Anxiety and 0.75 for Eating, respectively. Based on the posts in mental health forums on Reddit, Balani and de Choudhury [661] proposed an algorithm that can detect the levels of self-disclosure with an accuracy 0.78 using content features.

Literature have also leveraged SM data to detect depression. Moreno et al. [662] modeled the association between demographics and depression disclosures on Facebook. They found that 25% of Facebook user profiles display depressive symptoms, and 2.5% of them meet the criteria for a major depressive episode. Park et al. [663] identified features related to depression by analyzing data of 55 Facebook student users in Korea. They found that the response of users to tips has a positive correlation ( $r = 0.278$ ) with the CES-D scale (the ground-truth depressive symptomatology [664]), while the correlation ( $r = -0.237$ ) is negative for the number of friends. De Choudhury et al. [665] developed models to predict a mother's onset of post-partum depression (PPD) based on the Facebook data shared by 165 new mothers. They applied stepwise logistic regressions to predict the likelihood of PPD during the postpartum period using features extracted from the prenatal period alone. The model utilizing all features provides the most explanatory power, and the prediction accuracy can be improved by including the information in the early postnatal phase. The model using information of both the prenatal period and the early postnatal phase can explain about 48% of the variance in the data.

After compiling a set of 476 depressed users on Twitter, de Choudhury et al. [666] extracted behavioral features from their tweets over a year before the onset of depression. They found that some useful signals such as the decrease in social activity can characterize the onset of depression for individuals. Further, they built an SVM classifier with a RBF kernel [667] that can predict depression with an average accuracy about 0.70. Based on Weibo data, Wang et al. [668] developed a depression detection model by considering features of users only. They calculated the depression inclination of each post using a sentiment analysis method and constructed the detection model using features of depressed users. Their classifiers can detect depression users with an accuracy about 0.80. Tsugawa et al. [669] explored the effectiveness of using Twitter activity to characterize the depression level. They found that depressed users can be predicted with 0.69 accuracy using features extracted from their activity histories on Twitter. Moreover, two months is the optimal length of observation data from Twitter to identify depression.

Reece and Danforth [670] identified markers of depression from photographs posted to Instagram by 166 individuals. They extracted different features from the photographs and determined the strength of each feature using the Bayesian logistic regression. They found that depressed individuals have a higher posting frequency, post more comments, prefer to post photos with less face count, and are more likely to use filters. Further, they developed a suite

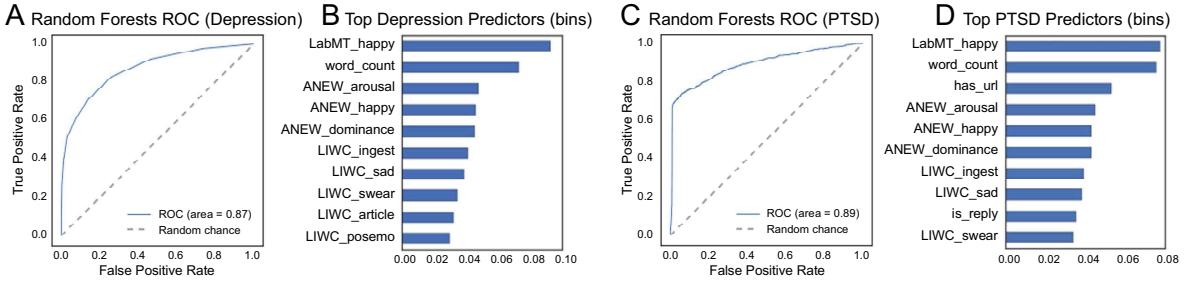


Figure 42: ROC curve and top predictors from the random forests algorithm, for depression (A and B) and PTSD (C and D) samples. Predictors with names ending with “\_happy” are happiness measures. LIWC predictors refer to the occurrence of semantic categories. Figure after [671].

of supervised machine learning algorithms that can achieve an F1 measure 0.647 in identifying depressed individuals. Reece et al. [671] predicted depression and PTSD users (see Figure 42) based on depression histories and Twitter behaviors of 204 individuals. After extracting several features from tweets, they trained supervised machine learning classifiers to identify depressed individuals. They found that a 1200-tree random forests classifier performs the best, where about 88.2% of PTSD predictions are correct. Further, they trained a hidden Markov model that can infer the onset of depression from tweets several months prior to clinical diagnosis.

Suicidality is a serious mental illness, and suicide attempt has extremely negative socioeconomic consequences. SM can be used as a surveillance tool to track suicidal behaviours. After analyzing Korean Naver blog data, Won et al. [672] found that dysphoria-related and suicide-related posts are predictive of suicide frequency under a multivariate model. In particular, dysphoria variables display a low variability and long secular trend, while suicide variables show a high variability and reaction to celebrity suicide events. Sueki [673] explored the relations between suicidal behaviours and suicide-related tweets using logistic regressions. They found that tweeting “want to commit suicide” is strongly related to suicide attempts. Abboute et al. [674] identified suicidal risky behaviors from tweets using language processing and classification methods. After collecting suspect tweets according to nine topics that suicidal people usually talk about [675], they classified into risky and non-risky tweets using six classifiers. They found that the naive Bayes performs the best in classifying risky tweets.

Tweets have been increasingly used to identify risky behaviors and track factors of suicide. Jashinsky et al. [676] found that southern and eastern states of the US have lower proportion of suicide-related tweets. The relevant tweets have a strong correlation (Spearman’s rank correlation  $\rho = 0.53$ ) with actual age-adjusted suicide data. Burnap et al. [677] evaluated a number of classifiers in classifying suicidal content and topics on Twitter. After extracting structural, lexical, emotional and sentiment features from tweets, they built an ensemble classifier based on the outcome of baseline classifiers. Their ensemble classifier can achieve an F1 measure 0.69 in classifying suicidal ideation. Benton et al. [678] proposed a multitask learning approach using a neural architecture to predict mental conditions based on Twitter data. Their approach improves all baselines models by giving an AUC about 0.84 at the best case.

Internet search queries have also been used to estimate suicide rates. Kristoufek et al. [679] explored the effectiveness of using Google Trends (GTs) to estimate suicide statistics in England. They found that larger search volumes of the term “suicide” indicate more suicides, while more searches for the term “depression” indicate fewer suicides. Moreover, suicide estimates based on GTs are much better than predictions based on previous suicide data. Tran et al. [680] validated GTs on predicting suicide rates in the US, Germany, Austria and Switzerland. They found that the associations between search volumes of suicide-related queries and suicide rates are weak. In particular, search volumes of the query “suicide” fail to show associations with suicide rates in the US. Thereby, they argued that the queries should be specific rather than broad in order to improve the performance of GTs-based suicide estimation.

Recently, Robinson et al. [681] reviewed literature that used SM for suicide identification and prevention. They summarized thirty articles published between 1991 and 2014 and concluded that to accurately assess suicide risk are still challenging. Mohr et al. [682] provided a review of mental health studies based on personal sensing, focusing on data from SM, smartphones, wearable devices, and so on. They provided a model for translating raw sensor data into behavioral markers that are related to mental health. Melia et al. [683] evaluated the effectiveness of interventions for suicide prevention based on mobile technology. They claimed that data-driven mental health analyses

have technological advances but ethical challenges.

## 5. Situational awareness and disaster management

### 5.1. Public health and epidemic surveillance

Public health surveillance is critical to social and economic systems since disease outbreaks may bring a huge burden on economics and a great damage on individual well-being. Scientists have shown that the origins of infectious diseases are significantly correlated with socioeconomic, environmental and ecological factors [684], and the human behavioural responses play an important role in epidemic spreading [685, 686]. A quantitative understanding of epidemic spreading is necessary for improving public health surveillance. To this point, a variety of models have been proposed to describe epidemic spreading on different types of networks [687, 688, 689, 690] and to understand the interplay between human mobility patterns and epidemic dynamics [691, 692]. Further studies have shown the possibility of locating the sources of diffusions with high credibility [693], for example, by designing a time-reversal backward spreading algorithm [694]. Recently, novel large-scale data and mathematical models have deepen our understanding of epidemic dynamics, helped to map epidemic activity [695] and suggested better immunization strategies to control epidemic spreading [696, 697, 698]. Indeed, the increasingly available data streams have been integrated into public health surveillance [699] and have contributed to the optimization of epidemic surveillance at multiple resolutions [700]. In this section, we will briefly introduce recent works that leveraged Internet search queries, social media (SM) data and mobile phone (MP) data to advance nearly real-time epidemic monitoring and surveillance.

#### 5.1.1. Search queries for epidemic surveillance

Real-time forecasts of influenza-like illness (ILI) outbreaks are usually hindered by the difficulties in collecting and analyzing a large volume of digital data in a timely manner. Traditional surveillance, relying on collections of clinicians' records and medical claims, is limited by data coverage, spatial resolution and long-time delay in delivering analysis results [700, 701]. Recent advances in information technology have made it possible to collect large-scale data of search queries that are related to public health. Useful health statistics regarding infectious disease activity can be yielded from health-related web searches. For example, Eysenbach [702] developed a method to analyze data collected from Google during the 2004/2005 flu season in Canada. They found that the number of clicks on links that triggered by entering “flu” or “flu symptoms” in Google is well correlated with the traditional disease surveillance data ( $r = 0.91$ ). Similarly, Polgreen et al. [703] found a correlation between the frequency of influenza-related searches in Yahoo! and the influenza activity. They built up linear models that can predict influenza increases three weeks in advance in the US. Pelat et al. [704] compared search trends of Google queries related to three infectious diseases in a French network. They found that search query data can also be utilized for infectious disease surveillance in a non-English-speaking country.

Google launched the Google Flu Trends (GFT) in 2008 as an Internet-based influenza surveillance tool, which uses aggregated Google search data to estimate ILI activity in real time. Specifically, Ginsberg et al. [705] analyzed Google search queries with influenza-like symptoms and proposed a method that can estimate the current level of weekly influenza activity in each US region with a reporting lag of about one day. Later, Cook et al. [706] evaluated the accuracy of the GFT model by comparing weekly estimates of ILI activity with the official data reported by the US ILINet. They found a high correlation ( $r \approx 0.94$ ) between the models' estimates (both the original and the updated GFT models) and the ILINet data, before and during the surveillance period. Yet, scientists have also pointed out some limitations of the original GFT model [707], for example, it predicted over double the proportion of doctor visits for ILI than the Centers for Disease Control and Prevention (CDC) did in 2013. Olson et al. [708] studied the reliability of GFT from 2003 to 2013 and argued that GFT may not provide reliable surveillance for seasonal or pandemic influenza. Changes in Internet search behavior and differences between the periods of GFT model-fitting and prospective usage diminish the performance of the original GFT model. Meanwhile, Lazer [709] suggested two issues that may contribute to GFT's drawbacks. One is the big data hubris as most data are not the output of instruments designed to ensure data's validity and reliability for scientific analysis. The other is the instability of the algorithm as changes made by engineers and consumers will affect the tracking of GFT.

The original GFT model has been well improved in recent years, and Google search queries have been widely used to nowcast influenza outbreaks. Based on real-time data with external information from GFT, Dugas et al. [710] developed a practical influenza forecast model that can provide accurate prediction of influenza cases. They developed the model by the generalized linear autoregressive moving average (GARMA) methods [711] with the negative binomial distribution. Formally, the GARMA( $p, q$ ) model is given by

$$\log(\mu_t) = X'_{t-1}\beta + \sum_{i=1}^p \phi_i [\log(y_{t-i}) - X'_{t-1-i}\beta] + \sum_{j=1}^q \theta_j \left[ \log\left(\frac{y_{t-j}}{\mu_{t-j}}\right) \right], \quad (81)$$

where  $\mu$  is the expected value of response ( $y$ ), and  $X$  is the vector of external variables with primes standing for transpose. The parameters  $\beta$ ,  $\phi$  and  $\theta$  of the model can be estimated from the training data. The model GARMA(3,0) can predict weekly influenza cases for 83% of the estimates with the out-of-sample verification, showing the capability of GFT in influenza forecasting. Using hospital influenza test results, Araz et al. [712] validated the usefulness of GFT in forecasting ILI-related emergency department visits. After testing five forecasting models, they found that linear regression models perform significantly better when including GFT data during 2008-2012.

Using Google searches between Jan 2010 and Sep 2013, Preis and Moat [713] built dynamic models to estimate the current level of influenza outbreak before the release of official data. They added GFT time series to an autoregressive integrated moving average (ARIMA) model [714] as an external regressor. The ARIMA model contains three autoregressive (AR) terms and two moving average (MA) terms (i.e., ARIMA(3,0,2)). Generally, the ARIMA( $p, d, q$ ) model is given by

$$y_{t+h} = \theta_0 + \sum_{i=1}^{p+d} \phi_i y_{t+h-i} + \sum_{j=1}^q \theta_j \varepsilon_{t+h-j} + \varepsilon_{t+h}, \quad (82)$$

where  $p$  is the number of AR terms,  $q$  is the order of the non-seasonal MA lags,  $d$  is the number of non-seasonal differences,  $h$  is the period in the future,  $y_t$  is the ILI level at week  $t$ ,  $\varepsilon_t$  is the white noise random error, and  $\phi_i$  and  $\theta_j$  are parameters to be estimated from data [715, 716]. They found that the in-sample forecasting mean absolute error (MAE) of the model is about 14% smaller than the baseline model without the GFT data. Teng et al. [717] developed a dynamic forecasting model to predict Zika Virus based on web searches from the Google Trends (GTs). They found a strong correlation between Zika-related GTs and the cumulative numbers of reported cases. Further, they constructed an ARIMA (0,1,3) model using online search data as the external regressor that can improve the prediction accuracy. Xu et al. [716] explored the predictive utility of Google search data in forecasting new ILI cases in Hong Kong by testing some individual models including generalized linear model (GLM) [155], ARIMA [714] and deep learning (DL) with feedforward neural networks (FNN) [20]. They found that DL with FNN are the best-performed algorithms in predicting the influenza peaks.

Data of Google search queries in other countries and from other platforms have also been used to improve infectious disease surveillance. Based on data of ILI-related Google search queries and historical CDC's ILI activity reports, Yang et al. [718] developed a methodological framework to produce retrospective estimates of ILI levels. Their multivariate linear regression modeling framework named ARGO (AutoRegression with GOogle search data) is expressed as

$$y_t = \mu_y + \sum_{j \in J} \alpha_j y_{t-j} + \sum_{k \in K} \beta_k X_{k,t} + \varepsilon_t, \quad (83)$$

where  $y_t = \log(c_t + 1)$  is the dengue case counts  $c_t$  at time  $t$ ,  $X_{k,t}$  is the Google search frequency of query term  $k$  at time  $t$ ,  $J$  is the set of auto-regressive lags,  $k$  is the set of Google query terms, and  $\varepsilon_t \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ . The ARGO model performs better in tracking ILI activity than some benchmark methods such as the GFT launched in 2014 (see Figure 43). Later, Yang et al. [719] extended ARGO to predict dengue cases in multiple countries/states, showing its nearly real-time ability to estimate dengue activity in data-poor environments. Yuan et al. [720] monitored influenza epidemics in China using Baidu search queries. They found a significant correlation between the selected composite keyword index and the case survey of Chinese influenza. Further, they fitted a linear model that can predict influenza cases one-month ahead for the first eight months of 2012 with the mean absolute error less than 11%. Li et al. [721] developed a dengue Baidu search index (DBSI), based on which they proposed a predictive model of dengue fever in Guangzhou, China. Their dengue early warning system combining DBSI with traditional surveillance and meteorological data can improve the capability of dengue case prediction.

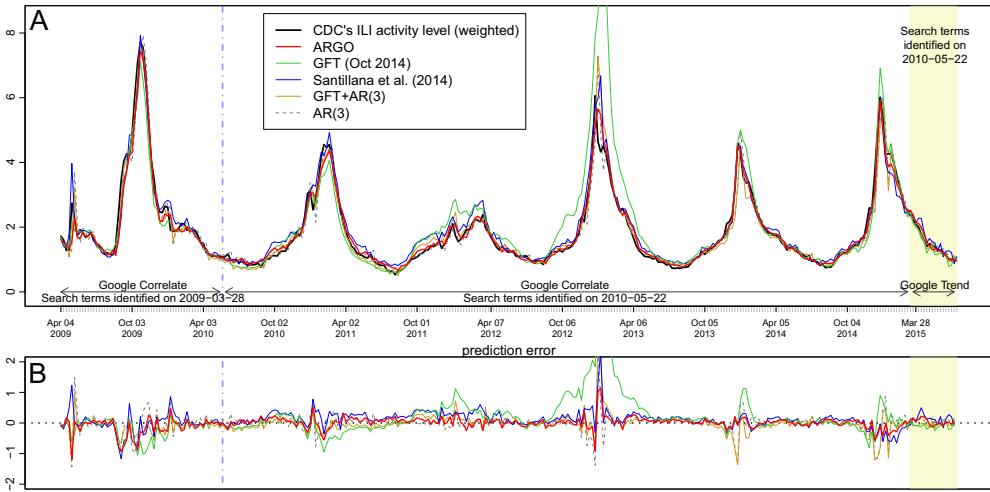


Figure 43: Comparison between estimations of ILI activity. (A) The estimated ILI activity level from ARGO (thick red), in comparison with the CDC's ILI activity level (black). The estimates from GFT (green), method of Santillana et al. [701] (blue), GFT plus AR(3) model (yellow), and AR(3) model (dashed gray) are also presented. The dash-dotted purple vertical line separates Google Correlate data into two periods. (B) The error of ILI activity estimation, defined as the estimated value minus the CDC's ILI activity level. Figure from [718].

### 5.1.2. Online posts for disease surveillance

Messages posted on Twitter are a new data source for disease surveillance. Chew and Eysenbach [722] analyzed over 2 million Twitter posts with keywords related to pandemic during the 2009 H1N1 outbreak. They found that Tweets containing “H1N1” increased from 8.8% to 40.5%, showing the potential of social media (SM) data on conducting infodemiology studies for public health. Culotta [723] identified influenza-related messages from over 500,000 messages spanning 10 weeks on Twitter. Then, they applied several regression models to examine the relations between the identified tweets and the CDC reported influenza statistics. They found that the best method exhibits a high correlation ( $r = 0.78$ ) with the CDC statistics. However, not all tweets containing influenza-related terms are suggestive to influenza outbreaks. For example, Aramaki et al. [724] found that 42% of tweets containing the word “influenza” are unrelated influenza tweets which do not refer to actual influenza outbreaks. Further, they developed an SVM-based classifier [725] to extract the mention of actual influenza patients from influenza-related tweets. Their method exhibits a very high correlation ( $r = 0.89$ ) in detecting actual influenza outbreaks.

Signorini et al. [726] collected a large sample of public tweets between April 29 and June 1, 2009 that contains pre-specified terms and the ground truth ILI data reported by the CDC. They trained a support vector regression (SVR) model [279, 280] on weekly term-frequency statistics. Their method produces estimates of national ILI values with an average error of 0.28%. Using 318,379 influenza-related tweets generated by 101,853 users, Salathé and Khandelwal [727] trained a machine learning algorithm to automatically judge sentiments of tweets. They found that sentiments expressed in tweets are positively correlated ( $r = 0.78$ ) with the official CDC-estimated vaccination rates at the regional level. Later, Lamb et al. [728] discriminated flu tweets that report actual infection from those that express concerned awareness of flu. Similarly, Broniatowski et al. [729] developed an influenza infection detection algorithm that can automatically identify relevant tweets. The estimates based on identified relevant tweets exhibits a strong correlation ( $r = 0.93$ ) with the CDC data, and their method can detect weekly deceleration of influenza prevalence with 85% accuracy.

Information embedded in Twitter stream has been shown very helpful in detecting rapidly evolving public concern with respect to the ILI emergence and transmission. Using over 287 million Korean tweets posted from October 2011 to September 2012, Kim et al. [730] developed regression models to track actual ILI epidemics and predict their activity levels. They maximized the correlation with the official reported ILI data by choosing a subset of markers and their weights using the LASSO regression method [731]. Their model has a significant improvement in prediction performance at the initial phase of ILI peak. Paul et al. [732] found that models incorporating influenza-related tweets can reduce the forecasting error by 17-30% compared with a baseline model using historical ILI data only. Moreover,

models relying on tweets perform better in estimating influenza prevalence than models using data from GFTs. Aslam et al. [733] analyzed 159,802 tweets collected from 11 US cities and found that tweets can serve as a supplementary surveillance tool for influenza with increased accuracy.

Topic analysis and topic models have been used to infer health concerns from Twitter data. Culotta [734] collected 4.31 million tweets of users in the US top-100 most populous counties and performed a linguistic analysis of geographical Twitter activity. They found a significant correlation on testing data for 6 of 27 health statistics, and they proposed models using Twitter-derived information, which can improve prediction accuracy for 20 of 27 health statistics. Chen et al. [735] proposed a temporal topic model to infer hidden biological states of users from their tweets. They developed an EM-based learning algorithm to model users' hidden epidemiological states. Their algorithm utilizing tweets from 15 countries in South America can learn meaningful word distributions and state transitions. Moreover, their algorithm can give better predictions of flu trends and flu peaks by aggregating states of users. Kagashe et al. [736] performed a topic analysis of widely used medicinal drugs during the 2012-2013 influenza season. They constructed an SVM classifier using dependency words as features to extract tweets that are suggestive to consumed drugs. Their model significantly outperforms some well-known benchmarks such as the lexicon-based model. Further, they extracted trending topics from drug-mentioning tweets using the LDA model [263] and found that the topic information of widely-consumed drugs can enhance seasonal influenza surveillance.

Researchers have developed several disease surveillance systems relying on data from Twitter and other SM platforms. Lee et al. [737] described a surveillance system to automatically predict seasonal disease outbreaks and monitor cancer activity levels based on tweets in the US. The resulted disease surveillance maps clearly show the distributions and timelines of disease types, symptoms and treatments. Meanwhile, Dredze et al. [738] presented a platform (HealthTweets.org) to share the latest surveillance results based on Twitter with public health officials. This platform provides three main visualizations including temporal health trends, specific locations and maps with a geographical view of health trends in the world. Indeed, health informatics datasets gathered from SM platforms have been increasingly applied to improve health care (see recent reviews [739, 740] and the references therein). To improve disease surveillance, Santillana et al. [740] suggested to leverage data from multiple sources such as online search, SM and traditional data. By combining ILI activity predictors of each data source, they developed an ensemble learning approach that outperforms GFT and autoregressive models by producing earlier estimates with a comparable accuracy.

Pageviews of disease-related Wikipedia articles have also been used to forecast seasonal influenza. By monitoring the rate of views on some specific Wikipedia articles, Mciver and Brownstein [741] developed a Poisson model that can accurately estimate ILI levels in the US in a timely manner. They collected Wikipedia article view data from December 2007 to August 2013 (294 weeks) and developed a generalized linear model to estimate ILI activity levels in the American population. Their model including 35 variables selected by the LASSO regression method can forecast the peaking weeks of ILI activity within a season more accurate than GFT. Later, Generous et al. [742] extended the above work for health purposes. In the same manner, they collected the Wikipedia article access logs from March 2010 to February 2014 and then applied a Poisson model fitted by the LASSO regression method to estimate ILI levels in the US. The estimates by their model exhibits up to  $R^2 = 0.92$  in predicting the official data with about one month in advance and with high feasibility across locations. Similarly, Hickmann et al. [743] leveraged Wikipedia access logs to create a weekly forecast for ILI activity levels during the 2013-2014 influenza season. Their linear regression model includes the weekly request data for influenza Wikipedia article and the previous week's ILI data as independent variables. They found that Wikipedia article access logs have a high correlation with historical ILI records (see Figure 44), and their method can accurately predict official reported ILI levels several weeks before the release.

The entire Wikipedia editing histories and article content are available [744], which provides a very rich data source for disease surveillance. Fairchild et al. [745] showed that Wikipedia content serves a centralized open-source monitoring and repository system, where disease-related data can be collected in real time. They used standard natural language processing techniques to identify key phrases in the content of disease-related Wikipedia articles such as death and hospitalization counts. Their method achieves an F1 of 0.753 in identifying relevant entities. Further, they analyzed articles of the 2014 West African Ebola virus disease epidemic and found that Wikipedia can provide detailed time series data, which are closely aligned with the data reported by WHO. Recently, Friedhorsky et al. [746] evaluated the use of Wikipedia access logs and category links for measuring global disease. They compared thousands of individual models for testing the effects of semantic article selection, forecast horizon, amount of training

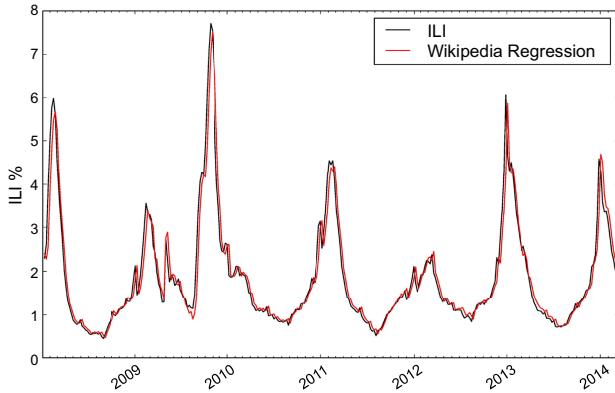


Figure 44: Regression of Wikipedia access logs to the officially released ILI data. The figure presents the linear regression of one week prior ILI observation, a constant term, and the access logs of five Wikipedia articles related to influenza to the current ILI observation. The regression of Wikipedia access logs is highly correlated to the true ILI outcome. Figure from [743].

data and model staleness based on Wikipedia data across six diseases and four countries. They found that the accuracy and robustness of disease estimation can be effectively improved by using minimal-configuration, language-agnostic article selection process based on semantic relatedness. Similarly, Sharpe et al. [747] provided a comparative analysis of Google, Twitter and Wikipedia data for influenza surveillance during the 2012-2015 influenza seasons. They detected seasonal change points by performing the Bayesian change point analysis [748] and calculated the sensitivity and positive predictive values (PPV) for each data source. They found that Wikipedia data have fewer change points in common with the CDC's ILI data, while they present the lowest sensitivity (33%) and PPV (40%) compared to Google and Twitter data.

### 5.1.3. Mobile phone records for epidemic prediction

Mobile phone (MP) data are a valuable source for studying the spreading dynamics of infectious diseases such as malaria, rubella, dengue, cholera and Chagas. The transmission of infectious diseases is significantly affected by human movements as infected individuals that make long-distance travels may transmit disease to healthy population in other regions. Therefore, quantifying human movements is critical to perceive and predict epidemic diffusion. Large-scale MP data have been leveraged to study human movements, which improves our ability of modeling and predicting the spreading of infectious diseases. Based on call detail records (CDRs) of 770,369 users for three months, Tatem et al. [749] estimated human travelling patterns and the imported malaria risk in Zanzibar. They extracted information on users' travelling and staying among Tanzania regions from CDRs, showing that imported malaria risks are heterogeneously distributed, where a very few people account for most of the malaria risk. Moreover, the likely sources and rates of malaria importation can be predicted by MP-derived human movement patterns in combination with the malaria endemicity data.

Based on data of 1.5 million Kenyan MP users, Wesolowski et al. [750] explored the dynamics of human carriers that drive malaria parasite spreading in Kenya. They mapped MP users to cell towers in settlements and assigned each settlement a malaria endemicity class according to a high resolution map of malaria prevalence. Then, they built travel networks of people and parasites between settlements and regions, focusing on the parasite importation by returning residents and by visitors from risky regions. Next, they examined directional and net movements of people and parasites between settlements by analyzing asymmetries between "source" and "sink" settlements. They found that the capital city Nairobi and its surroundings are a major destination for both humans and parasites. Moreover, returning residents play an important role in importing parasites to major parasite sinks, and some local transmission may occur in residential and less developed areas on the periphery of the city. Wesolowski et al. [751] further explored the CDRs data and found that human mobility measures extracted from MP data can predict which region is lacking preventive care. Tatem et al. [752] predicted malaria risk by analyzing CDRs of about 1.5 million users and case-data of risk maps in Namibia. They found that most of the northern Namibian areas are major sources and sinks of parasites, and the heterogeneity in human movement patterns results in the variation in the risk connectivity of sources

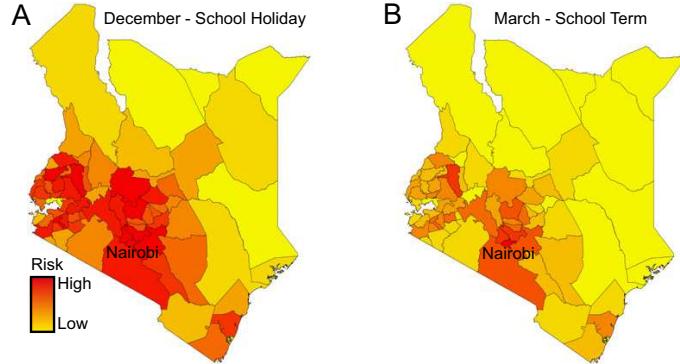


Figure 45: The seasonal variability in the risk of rubella importation in Kenya. There are large amounts of population flux and consequently the risk of rubella importation per district (A) during the major holiday and a school term break (December), while the rubella importation risk decreases (B) during the school term. The western Kenyan districts are with higher risks during the year, and the capital city Nairobi consistently remains at a high risk of importation. Figure from [756].

and sinks. Moreover, targeting control on malaria risk connectivity in certain areas with larger exporters (sources) of infections can largely affect their surrounding areas. These works highlight the value of MP and complimentary data on perceiving and controlling infectious disease.

Multiple data sources have been intergraded to better understand disease spreading and implement targeted surveillance. Wesolowski et al. [753] collected travel data of 2,650 individuals in two Kenyan districts through a malariometric survey and compared it with the data of about 1.5 million MP users [750]. They found that both survey and MP data can predict the amount of imported malaria and identify the major routes, however, the two datasets have a wide divergence in terms of the travel magnitude. For example, travel surveys can provide information on demographics, travel motivations and destinations [754], but they tend to under-estimate the volume and range of human mobility. Although the distributions of human movements estimated based on surveys and MPs are highly correlated with each other, Tizzoni et al. [755] observed a systematic overestimation of commuting traffic in MP data. That is because MP-sampled population can not perfectly represent the general population. In addition, human mobility proxies perform differently in approximating commuting patterns for disease spread at different resolutions, suggesting that the chosen of mobility proxies should account for the epidemic situation under study. Regarding this point, Wesolowski et al. [756] showed that MP-derived seasonal and spatial travel patterns are predictive for disease transmission. They produced dynamic importation risk maps for rubella in Kenya (Figure 45) using MP data. Their work demonstrates the effectiveness of MP data on identifying critical drivers of epidemics on relevant spatial and temporal scales.

In-depth analyses of human mobility patterns help improve the prediction of geographic spread and timing of epidemics. Based on data of about 40 million MP users, Wesolowski et al. [757] quantified human travel patterns underlying the spread of dengue virus in Pakistan. They estimated the travels of users based on the MP data and then fitted an ento-epidemiological model [758] to the reported dengue cases in southern Pakistan. They estimated the timing of the first imported case in northern Pakistan by employing an epidemiological approach, and they found that MP-based mobility estimates can predict the geographic spread and timing of epidemics. Further, they generated fine-scale dynamic risk maps by combining estimates of seasonal dengue importation and transmission suitability maps. Formally, the epidemic risk for a location  $x$  is given by

$$\text{risk}(x) = \sum_{t=1}^N Z_x(T_t) Y_{x,t}, \quad (84)$$

where  $Z_x(T_t)$  is the environmental suitability for dengue, and  $Y_{x,t}$  is the importation of infected travelers on day  $t$ . Based on MP data of about 2.9 million users in Haiti, Bengtsson et al. [759] explored the influence of population mobility on the spatial evolution of a large-scale cholera outbreak. They predicted the risk of a new outbreak in an area by calculating the infectious pressure based on the human mobility. They found that the infectious pressure at outbreak onset is linearly correlated ( $r \approx 0.3$ ) with the average daily number of reported cases within seven days of the new outbreak. Finger et al. [760] collected MP data of about 150,000 users to estimate human mobility fluxes

during the 2005 Senegal cholera outbreak period. They developed a mechanistic model that takes into account human mobility and other drivers of cholera disease transmission such as rainfall. They found that mass gatherings during the initial phase of the outbreak can significantly affect the spreading of waterborne diseases cholera.

Human mobility patterns derived from MP data have been introduced into the mathematical modeling of infectious disease transmission. By analyzing the MP data [761] released for the Data for Development (D4D) project, Tompkins and McCreesh [762] identified the characteristics of human movements involving overnight stays, which are relevant for malaria transmission. They found that about 60% of people have regular destinations that they visit repeatedly, and the number of overnight journeys peaks at a distance of 50 km. Further, they proposed an agent-based migration model by adapting a gravity model to describe overnight journeys. Their model can well reproduce general population mobility patterns driving malaria transmission. Based on CDRs during the chagas spreading in Argentina and Mexico, de Monasterio et al. [763] analyzed mobility patterns of users and predicted the movements among different regions. They detected possible risk zones of chagas disease and produced the risk maps for two Latin American countries. Recently, Wesolowski et al. [764] explored how seasonal variation in human movement affects infectious disease dynamics based on MP data collected from Kenya, Namibia and Pakistan. They found that major national holidays will lead to seasonal fluctuations in human mobility, which further result in seasonal fluctuations of the country-scale connectivity. Using a spatial diffusion model (see Ref. [764]), they evaluated the consequences of directional asymmetries and seasonal variation in travels as well as pathogen characteristics on epidemic spreading, and found that the spreading speed depends not only on the pathogen's characteristics but also on the month that the pathogen is imported.

Panigutti et al. [765] built two human mobility networks in France, namely, the MP commuting network and the census commuting network. Then, they compared 658,000 simulated epidemic outbreaks generated using a reaction-diffusion (RD) metapopulation model based on the two mobility networks. The RD dynamics are separated into two components, home time and work time. The number of susceptible, infected and recovered individuals [766] who live in district  $i$  and work in district  $j$  are respectively defined as  $S_{ij}$ ,  $I_{ij}$  and  $R_{ij}$ . The spreading rate is  $\beta$ . The force of infection during home time  $\lambda_i^{\text{home}}$  is defined as

$$\lambda_i^{\text{home}} = \beta \frac{I_{ii} + \sum_{j \in \nu_i} I_{ij}}{N_{ii} + \sum_{j \in \nu_i} N_{ij}}, \quad (85)$$

where the sums run over the neighbourhood of district  $i$ :  $j \in \nu_i$ . The force of infection during work time  $\lambda_i^{\text{work}}$  is defined as

$$\lambda_i^{\text{work}} = \beta \frac{I_{ii} + \sum_{j \in \nu_i} I_{ji}}{N_{ii} + \sum_{j \in \nu_i} N_{ji}}, \quad (86)$$

where  $N_{ij} = S_{ij} + I_{ij} + R_{ij}$  is the total number of commuters living in district  $i$  and working in district  $j$ , and  $N_{ii}$  is the number of residents in district  $i$  who also work in district  $i$ . The numbers of susceptible individuals in district  $i$  during home and work time are respectively given by  $S_i^{\text{home}} = S_{ii} + \sum_{j \in \nu_i} S_{ij}$  and  $S_i^{\text{work}} = S_{ii} + \sum_{j \in \nu_i} S_{ij}$  [765]. Simulation results show that MP data are more reliable in describing human movements in central regions. Moreover, it is essentially important to obtain an accurate estimation of epidemiologically relevant mobility patterns in the seed area in order to capture future spreading patterns of the outbreak. Mari et al. [767] explored the drivers of endemic schistosomiasis by parameterizing a spatially explicit network model based on a large dataset of MP traces. They found that the epidemic prevalence may be reduced by moderate mobility while increased by either high or low mobility. Moreover, environmental and socioeconomic heterogeneities play a crucial role in capturing the spatial epidemic prevalence, and the inclusion of human mobility patterns can significantly improve the predictive power for the infectious disease spreading.

There are opportunities yet challenges of leveraging MP data to link human mobility patterns with infectious disease dynamics. Wesolowski et al. [768] pointed out some limits such as the availability of MP data, the biases on MP ownership, and the lack of large-scale demographic or social identifiers. Meanwhile, they suggested some opportunities including the fine-scale individual movements across large numbers of individuals, the new data of GPS for understanding social connections, and the data access for public health interventions. In a broader perspective, Jones et al. [769] reviewed some challenges and potential opportunities of using CDRs for public health research. They pointed out that the majority of previous studies paid attentions to below middle-income countries, CDRs were mainly used in aggregated form to estimate population movement, while public views on using CDRs for public

health research were lacking. To address these issues, they suggested to develop an ethically founded framework to gain public views and to better integrate routine health records with validated CDRs in future public health research.

## 5.2. Emergency and disaster monitoring

Along with increased urbanization and changing climate, many areas are now facing an unprecedented number of emergent events and natural disasters, which pose numerous threats to human life and economic development. It urges rapid situational awareness and efficient management strategies to reduce human suffering and economic losses [770, 771]. In rural areas, assessments of natural hazards usually follow a delay, resulting in difficulties of disaster response and relief. In urban areas, detections of emergent events (such as terrorist attacks, riots and large-scale demonstrations) and natural disasters (such as earthquakes, floods and hurricanes) are critical not only for governments' rapid disaster response [772] but also for in-depth understanding of human behaviors in extreme situations that will further help in better designing strategies in disaster relief [773]. In addition to theoretical methods [774, 775], novel data sources have been leveraged to improve emergency awareness and disaster management such as remote sensing (RS) [776], mobile phone (MP) [777], and social media (SM) [778], with remarkable advantages of low acquisition cost, real-time updates and high spatio-temporal resolutions.

### 5.2.1. Remote sensing for disaster assessment

Mapping natural hazard and disasters is critical for emergence response, disaster relief and crisis-management support [779]. However, disaster management relying on site surveys and field observations usually requires many resources, follows a long-time delay and is constrained by time and space. Fortunately, these problems can be tackled by using the increasingly available RS data as they update timely with low cost, have high spatial and temporal resolutions and capture a wide field of view [776]. In recent years, a number of pioneering works have illustrated the utilization of RS data in combination with image processing techniques for a rapid damage assessment of natural disasters such as earthquakes, flooding, wildfire and landslides [780]. The RS-based disaster management is necessary for supporting, for example, damage assessment and relief priority map. In the following, we will introduce some applications of RS data for earthquake damage assessment and flood monitoring.

High-resolution satellite images can be used to detect changes of ground surface and buildings before and after earthquake. This opens new opportunities for earthquake damage assessment at the level of settlements and buildings. Using synthetic aperture radar (SAR) interferometry images obtained by the ERS-1 satellite, Massonnet et al. [781] captured the ground surface movements caused by the 1992 Landers earthquake in California, which agree well with surveying measured displacements. Miura and Midorikawa [782] detected locations of newly constructed buildings from high-resolution satellite images and updated GIS building inventory data. They conducted the building damage assessment for a scenario earthquake in Metro Manila, Philippines. Marin et al. [783] developed an approach to detect building changes before and after earthquakes from very high resolution (VHR) SAR images. They extracted information on changes by analyzing the exploitation of expected backscattering properties of buildings. Validated using spotlight images of two Italian cities, their approach shows a high reliability in identifying demolished buildings.

Satellite images from RS have been used to map earthquake exposure and conduct damage assessment after the 2010 Haiti Earthquake. Based on pre- and post-event VHR satellite imagery, Corbane et al. [784] produced a building damage assessment map for the 2010 Haiti Earthquake. They compared the reliability of this area-based map to the detailed damage assessment derived from the post-event aerial imagery. Result suggests that satellite-based damage assessment maps are able to capture the damage pattern especially in heavily damaged areas, however, they cannot provide sufficient information to quantify damage intensity. Uprety and Yamazaki [785] detected buildings that were damaged during the 2010 Haiti Earthquake by calculating the backscattering difference (as well as correlation) between two SAR images taken before and after the earthquake. Later, Ehrlich et al. [786] showed that building damages can be detected from pre- and post-disaster VHR imagery, and the measured damages can provide vulnerability information related to the structural fragility of building stocks. Tian et al. [787] developed a novel method to monitor after-disaster building damages. Two post-event satellite stereo imageries were combined with digital surface models, panchromatic images were segmented, and a rule-based classification was used to identify collapsed buildings.

Satellite imagery has also been used to locate affected areas after large earthquakes in China. One of the most severe natural disasters in China is the 12 May 2008 Wenchuan Earthquake, which changed the entire landscape of the affected area. Based on VHR satellite images, Liou et al. [788] investigated landslides and their consequences

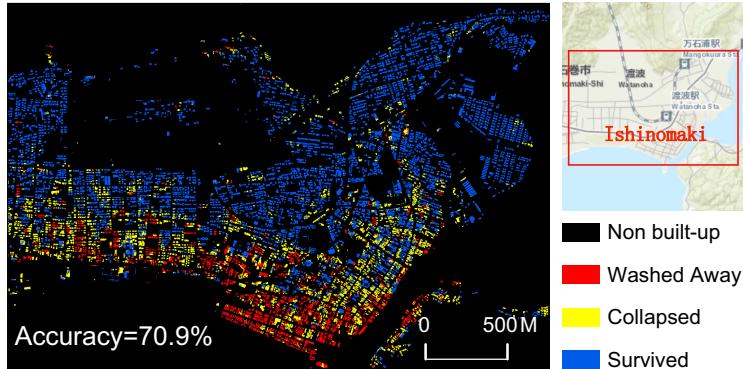


Figure 46: Results of the damage map using the U-net convolutional network. The south-eastern part of Ishinomaki city is used as the validation area during the 2011 Tohoku Earthquake-Tsunami. The figure presents the survived areas (blue), contrasting with collapsed areas (yellow), washed away areas (red), and not built-up areas (black), identified by the U-net convolutional network. The overall accuracy of damage classification is 0.709. Figure from [797].

following the Wenchuan earthquake. They identified structural deformation of land areas and rupture of dams and suggested some precautionary measures to avoid further destruction. Tong et al. [789] proposed an approach to detect collapsed buildings based on satellite stereo image pairs taken before and after the Wenchuan earthquake. They detected an individual collapsed building by the height differences and identified the region of collapsed buildings by differentiating the digital elevation models generated from the image pairs. Their method can accurately estimate the status of collapsed storeys and determine the collapsed region with an overall accuracy of over 90% assessed using an error matrix [790]. The 2010 Yushu Earthquake is another large earthquake causing many buildings collapsed in China. For the Yushu earthquake, Jin et al. [791] derived a building damage assessment from VHR SAR images. After analyzing the features of collapsed, partial collapsed and non-collapsed buildings in the SAR images, they counted the number of buildings with different damage levels and built the damage index for each block. Shi et al. [792] extracted multiconfiguration features to distinguish the fallen and intact structures. They employed a random-forest framework to quantify the importance of each feature to improve the accuracy.

Recently, deep learning algorithms have been introduced to analyze RS data for rapid earthquake damage mapping [20, 793]. For the 2010 Haiti Earthquake, Cooner et al. [794] evaluated the effectiveness of several deep learning algorithms in detecting earthquake damage. They found that spatial texture and structure features extracted from satellite images are more important than spectral information in classification. Multilayer feedforward neural network can detect damaged buildings with an error rate below 40%. By combining convolution neural networks (CNNs) and multiscale segmentation, Sun et al. [795] proposed a method to map earthquake damage from VHR images. They firstly trained CNNs to get initial classification about original images, and then combine the initial classification with the results of multiscale segmentation, to obtain class-based segmented images with different scales. Their approach performs very well in mapping the Wenchuan earthquake damage. Fujita et al. [796] applied CNNs to detect damage buildings from pairs of satellite image patches taken before and after the 2011 Tohoku Earthquake-Tsunami. Their CNN-based detection system can classify washed-away buildings with accuracy in the range 94-96%. Recently, Bai et al. [797] developed a deep learning algorithm to map damage, where the U-net convolutional network is employed to semantically segment VHR satellite images. Their algorithm can classify damage with an overall accuracy 0.709 based on pre- and post-disaster images of the Tohoku tsunami (see Figure 46). In addition, the damage map can be updated in every 2-15 minutes when images are available.

RS data play an indispensable role in operational practice of earthquake response practice. Through analyzing high-resolution satellite imagery, not only public health needs can be rapidly assessed, but also relief priority map can be suggested. For some isolated earthquake areas with uneven damage distribution, RS-based damage assessment can promote the effectiveness of rescue efforts. By leveraging GIS technology and high-resolution satellite images, Zhao et al. [798] developed an assessment framework to rapidly evaluate health loss. Their method can estimate casualties and injuries with an accuracy 0.77 within a few hours for the 2008 Wenchuan earthquake. Moreover, they can identify damaged medical institutions, mark high-risk areas of schistosomiasis and map temporary settlements for victims.

Hamid et al. [799] determined the degree of building damage based on the texture features extracted from pre- and post-event VHR satellite imagery. They proposed an algorithm that can produce relief priority maps after earthquakes. For the Varzaghan earthquake in Iran, their method achieves a general accuracy 0.88 in classifying damaged buildings into three classes of negligible damage, substantial damage and heavy damage with relief priorities from low to high. Their relief priority maps can guide rescue teams under limited resources after earthquakes. Recently, Quinn et al. [800] reviewed applications of machine learning approaches for refugee settlement mapping based on RS data. The combination of machine learning algorithms and RS data provides a way to better coordinate humanitarian relief.

Flood is one of the most serious disasters which can destroy homes, cause mudslides and take human lives. Flood management requires timely awareness of flood situation, locating flooded areas and implementing damage relief [801]. Satellite- and aircraft-based RS can provide the required information with high spatial and temporal resolutions. In recent years, RS has become a useful tool in flood management [802]. By integrating satellite images with ancillary information from GIS, Brivio et al. [803] proposed a procedure to estimate flooded areas at the peak time. For the 1994 flood in northern Italy, their method can identify 96.7% of flooded areas compared with the official reference map. Groeve [804] described a method to early detect floods from satellite imagery. For the 2009 Southern Africa flood and the 2010 Namibia flood, their method can provide early flood warning up to 30 days by monitoring upstream areas and detect floods two hours after their occurrences. Skakun et al. [805] assessed flood risk based on time series of satellite images. They calculated the relative frequency of inundation (RFI) based on the flooded areas extracted from satellite images and the maximum flood extent images produced for previous flood events. The RFI map serves as a hazard map for flood risk assessment. Their method can identify cities and villages with the highest flood risk for the Namibia flood.

RS technology can produce high-resolution flood hazard maps for regions lacking of ground based system to monitor rainfall and river discharge. Giustarini et al. [806] combined multiple annual satellite observations and continuous spatially-distributed hydrodynamic model to map flood hazard with a high spatial resolution. According to the study on the UK Severn River flood, their method exhibits advantages in high-resolution flood mapping against the reference map computed by the hydraulic modeling approach. Moreover, their method has merit on the flexibility, where any type of RS images can be included as the inputs. Kwak [807] derived the so-called synchronized floodwater index (SfWi) from annual time-series optical satellite data to detect the maximum extent of a nationwide flood. For the 2015 monsoon season, they revealed the propensity of flood risk in three major rivers by analyzing the spatio-temporal dynamics of the maximum flood extent. Flood areas suggested by SfWi are small but accurate. Rahman and Di [808] reviewed the applications of the state-of-the-art RS techniques for flood management. On flood risk assessment, RS can be used to assess flood risk, exposure and vulnerability. On flood emergency planning, RS has contributed to flood warning system, rescue and relief operation, post flood damage assessment, and policy making.

The production of high-precision flood maps requires the integration and classification of information coming from different RS data sources. D'Addabbo et al. [809] applied Bayesian networks to monitor flood based on multi-temporal and multi-sensor RS data. They used Bayesian networks to perform a data fusion procedure of different types of satellite imagery and ancillary data. Other open-access data can also supplement commercial satellite imagery and ground-based data in developing regions where relevant data are usually sparse. Ekeu-wei and Blackburn [810] introduced applications of open-access RS data (such as altimetry, DEMs, optical and radar images) on mapping flood. Using Nigeria as a case study, they evaluated the significance of open-access datasets in flood risk assessment and found that open-source data provided by the private sector play a key role in assessing flood risk especially for data sparse regions. Readers are encouraged to read a recent book by Refice et al. [811], who reviewed the state-of-the-art RS techniques and useful tools for flood hazard monitoring.

### 5.2.2. Mobile phones for emergency management

Mobile phone (MP) is a useful tool not only to facilitate the daily communications between users but also to record their geographic positions with high spatial and temporal resolutions. As shown in the previous sections, MP data is powerful in inferring individual socioeconomic status, analyzing offline human mobility and exploring online activity patterns. However, these quantitative features of human activities are usually under normal and stationary circumstances. Yet, human behavioral patterns are intuitively different under unfamiliar situations, especially during emergent events. Therefore, human activity patterns revealed by MP data have promising applications on detecting, monitoring and managing emergent events [812]. On the one hand, analyzing information flow recorded by MPs can help detect emergent events in real time. People under extreme circumstances will change their calling and behavioral

patterns (see Ref. [813] for calling and behavioral patterns under normal circumstances). Dramatic behavioral changes can be treated as signals of emergent events, with high-resolution time and locations. On the other hand, MP data can help provide real-time emergency monitoring by analyzing mobility patterns. In a word, MP data can be used to quantify human emergency behavior and track population evacuation.

Large-scale MP datasets have potential applicability in real-time situation awareness and emergency detection. Bagrow et al. [814] explored human societal response to external perturbations based on a country-wide dataset of MP communications covering about ten million users. They compared the real-time changes in mobile communication patterns between eight emergencies and eight non-emergencies, and found that calling activities under emergencies spike rapidly and decay immediately after the event, and the call volume decays exponentially with the spatial distance. Moreover, people affected by emergencies propagate the emergency information rapidly and globally through social networks. Moumni et al. [815] explored social response to the 2012 Oaxaca earthquake in Mexico based on call detail records (CDRs) for two weeks. They analyzed four different variables including call volume, call duration, social activity and mobility, showing three stages of the social response: a spike of many short calls in five minutes after the earthquake, a reduced activity of short calls lasting one to two hours and a moderate increase in call and duration volumes lasting about 5 hours. Moreover, users tend to moderately increase mobility during the earthquake.

MP data can be used to explore collective call behaviors and information flow following emergencies. Based on MP activities of about 10 million users in an European country, Gao et al. [816] studied the information spreading and the changes of users' communication patterns during emergencies. In consistent with the previous findings [814, 815], they observed a sharp increase in call volume during an emergent event. To explain the volume spikes, they analyzed reciprocal communications by decomposing them into call-forward and call-back. They found that call-back response and dissemination of emergency information have an effect on the magnitude of volume spikes but the former is the dominant component. Moreover, the observed reciprocal communications, in particular, the call-back response during emergent events, can not be explained by the inherent reciprocity in social networks under normal circumstance. Taking three days' CDRs during the 2012 Xinjiang earthquake in China, Yu et al. [817] analyzed the collective call patterns of people who experienced the earthquake. They found that earthquake significantly increased many indices of call patterns such as call volume and call duration. In particular, call volume gets increased more significantly. From a spatial perspective, people made more local calls on the earthquake day. From a temporal perspective, local call volume raised rapidly within two hours after the earthquake, and large volume of distant calls last a whole day.

Calling behavior changes during flooding events can be tracked by MPs. By analyzing CDR data and RS data of the 2009 Tabasco floods in Mexico, Pastor-Escuredo et al. [818] revealed abnormal human communication patterns during and after the events. They identified the floods from satellite images and reconstructed a flood impact map. To detect abnormalities in the communication activity from the CDR data, they calculated the variation metric at the cell tower by comparing the MP activity  $x(t)$  during the disaster against their characteristic variation obtained during the baseline (BL) period. Formally, the variation metric  $x_{\text{norm}}$  is defined by

$$x_{\text{norm}} = \frac{x(t) - \mu_{\text{BL}}}{\sigma_{\text{BL}}}, \quad (87)$$

where  $\mu_{\text{BL}}$  and  $\sigma_{\text{BL}}$  are respectively the mean value and standard deviation of the activity during the baseline period. They found that the variation metric spikes in most flooded areas and shows consistence to the flood impact map, suggesting that the CDR-based variation metric can be used for situation awareness. Based on CDRs of Senegal, Hong et al. [819] explored how different levels of floods affect MP call volumes and communication network features. They found that people increase their calling activities during floods, and the call volumes are positively correlated with the flooding intensity. Moreover, large cities with more recurring floods have more introversion communications within their neighborhoods, suggesting that people living in larger cities rely more on local communities.

Rapid emergency detection based on MP data can facilitate humanitarian response and reduce the toll of extreme events. Based on the combined data of MP activities and official event records in Rwanda, Dobra et al. [820] proposed an efficient system that can detect days with anomalous behavioral patterns under many emergent and non-emergent events. They found that days with increased anomalous behaviors suggest joyous events, while days with decreased anomalous behaviors suggest emergent events. Thus, the type of events can be identified by examining the increases and decreases in anomalous behaviors. Moreover, they confirmed that the behavioral responses have significant spatial and temporal variances. Recently, Gundogdu et al. [821] proposed an approach based on Markov modulated Poisson process to detect behavioral anomalies from CDRs of 5 million users. They found that different types of events have

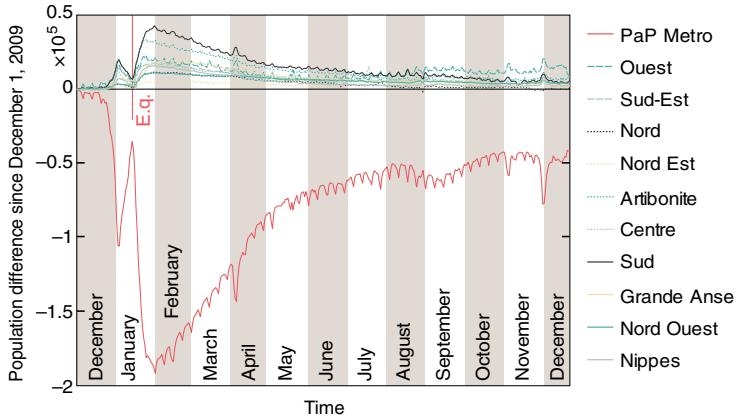


Figure 47: Population movements following the 2010 Haiti earthquake. The figure presents the changes in the number of individuals in the various provinces (color coded curves) before, during and after the earthquake. The red vertical line marks the date of the earthquake. The red curve highlights the capital PaP Metro. Figure from [823].

spatial and temporal differences in call volume changes, and emergent events can be distinguished from non-emergent events by comparing their associated calling patterns. Moreover, communication activity is more significant than mass movement in event detection.

In addition to analyzing communication behaviors, MP data have been increasingly used to model human mobility perturbations, assess population displacements and improve emergency responses during large-scale disasters. The subscriber identify module (SIM) cards are useful in tracking human movements in real time. Using records of 1.9 million SIM cards before and after the 2010 Haiti earthquake, Bengtsson et al. [822] estimated the trends and magnitudes of population movements after the January earthquake and during the October cholera outbreak. They found that SIM cards could provide valid estimates of the distribution, magnitude and trends in population displacements. In particular, the geographic distribution of the estimated population moving out of the capital city Port-au-Prince (PaP) is highly heterogeneous and consistent to the retrospective survey data, while the destination of population movements out of the cholera outbreak areas is heavily concentrated. Based on the same dataset, Lu et al. [823] explored the predictability of population displacements after the Haiti earthquake. They found that the population in PaP decreases by 23% in the three-month period after the earthquake due to population movements (see Figure 47). Although people's mobility patterns sharply changed with a growth in travel distances and trajectory sizes, the predictability of population movements remains high. In particular, they found a high correlation between the destinations of people who left PaP during the first three weeks and their mobility patterns during normal times. People tend to move to locations where they have significant social bonds. The results suggest a high-level predictability of human spatial regularities even under extreme events [824], which will help predict disaster responses and manage population movements.

Accurate prediction of human emergency behaviors is critical to disaster management and societal reconstruction. Song et al. [825] constructed an enormous set of GPS mobile sensor data recording about 1.6 million people's activities during the 2011 Japan Earthquake and the Fukushima nuclear accident. They revealed the short-term and long-term evacuation behaviors after the disasters. For example, the population in affected areas substantially decreased by more than 50% during the first 12 days and stabilized 81 days after the earthquake. Further, they trained a general probabilistic model to automatically predict population evacuations in affected cities. Their method can find new mobility features and predict large population movements with an accuracy about 0.80. Song et al. [826] further found that human emergency mobility after disasters is correlated with their mobility patterns during normal times and highly affected by many factors including the intensity of disaster and social relationships. Taking into account the above factors, they developed a model that performs better than some well-known baseline models on predicting human emergency mobilities. The model outperforms their previous method [825] by up to 11.08% in human mobility prediction. These results suggest that human mobility under extreme events is more predictable than our intuition.

MP data have also been combined with other data sources to analyze human behavioral changes and assess population displacement during emergencies. Based on MP records and satellite images, Bharti et al. [827] analyzed

population movements during the 2010 Côte d'Ivoire internal political conflict. The two datasets strongly agree with each other on estimating average population sizes pre- and post-conflict, and they complement each other on estimating long- and short-term population dynamics throughout the crisis. Based on CDRs of 12 million MP users, Wilson et al. [828] explored the rapid assessments of the national-level population displacements after the 2015 Nepal Earthquake. They uncovered the patterns of return to earthquake affected areas. An estimated number of 390,000 people in earthquake affected areas evacuated immediately to surrounding areas, in particular to those with high populations. Most people will gradually return to their hometown after the earthquake, while less than 15% people were still away from their home three months after the earthquake. Ghurye et al. [829] analyzed the changes of mobility patterns and communication behaviors based on CDRs during the 2012 Rwanda flood season and found that disasters disrupted mobility patterns. During the first three weeks, the number of victims who left their hometown reaches its peak at about 10 millions, and the recovery to normal patterns takes over two months.

### 5.2.3. Social media for situational awareness

When people facing extraordinary events, their collective attentions and behaviors will emerge on social media (SM). By tracking Twitter users' hashtags, He and Lin [830] quantified the shift of human collective attentions under exogenous shocks. They found that the co-occurrence network of users' hashtags exhibits a strong community structure before the event, while a few hashtags will suddenly appear in many tweets and thus become hubs after the event. Sano et al. [831] analyzed the keyword appearance rate based on more than 1.8 billion Japanese blog entries. They found that the functional forms of decay and growth of keyword appearance that peaked on a certain day exhibit power laws with the various exponents values between  $-0.1$  and  $-2.5$ . In particular, the absolute exponent value is less than one for some keywords of news during extraordinary events.

SM is a valuable source of information for gaining situational awareness, detecting and locating emergent events, improving disaster response and enhancing relief efforts. Yet, the extremely high volume of messages generated during crises urges for automatic methods that can extract relevant and valuable disaster-related information from SM posts [832]. For example, Kireyev et al. [833] collected two datasets from Twitter during earthquakes and applied topic models to analyze earthquake-related tweets. Imran et al. [834] utilized the state-of-the-art machine learning techniques to extract disaster-related information from tweets generated during the 2011 Joplin tornado. Based on a large set of tweets, Olteanu et al. [835] created a lexicon consisting of crisis-related terms that frequently appear in messages posted during various crisis situations. Such a crisis lexicon can be used to improve the recall but maintain the precision in the sampling of crisis-related tweets. Further, they showed how to automatically identify the terms describing a given crisis based on the crisis lexicon. For disaster response and relief, Ashktorab et al. [836] proposed a Twitter-mining tool named Tweedr that can rapidly extract relevant information from tweets posted during disasters. The Tweedr pipeline has three phases, where disaster-related tweets are identified in the classification phase, similar tweets are merged in the clustering phase, and tokens and phrases of damage information are extracted in the extraction phase. The Tweedr can identify 12 crises events occurred in the US since 2006.

The utilization of SM data has transformed the methodology of earthquake detection and early warning [837], where the distribution of shakings can be mapped in minutes from earthquake-related posts. Acar et al. [838] studied earthquake information sharing on Twitter by analyzing the tweets posted near two disaster-struck areas during the 2011 Tohoku Earthquake. They found that people in directly affected areas tweeted to announce their uncertain and unsafe situation, while people in remote areas tweeted to inform followers that they are safe. Toriumi et al. [839] analyzed 360 million pre- and post-disaster tweets for the 2011 Tohoku Earthquake. They found that users changed their main purpose of using Twitter from communication to information sharing after the disaster. In particular, critical information was widely retweeted while non-emergency tweets were decreased. For the 2012 Indonesia Earthquake, Chatfield and Brajawidagda [840] identified 6,383 earthquake-related tweets and performed a social network analysis of the information flows on Twitter. They showed that Twitter can be utilized as an early warning network, where the followers of governmental agencies will retweet warnings immediately. Dong et al. [841] analyzed information diffusion on Weibo after two earthquakes in China. They found that strangers play an important role in spreading earthquake-related news, and verified users dominantly influence information diffusion on Weibo.

People post many earthquake-related messages on SM soon after they feel shaking. Social media users indeed serve as social sensors with their posts being the sensory information. Sakaki et al. [842] proposed a method to detect target events by leveraging the real-time and geographical nature of Twitter. They employed a support vector machine (SVM) to classify tweets related to target events and proposed a probabilistic spatio-temporal model for each

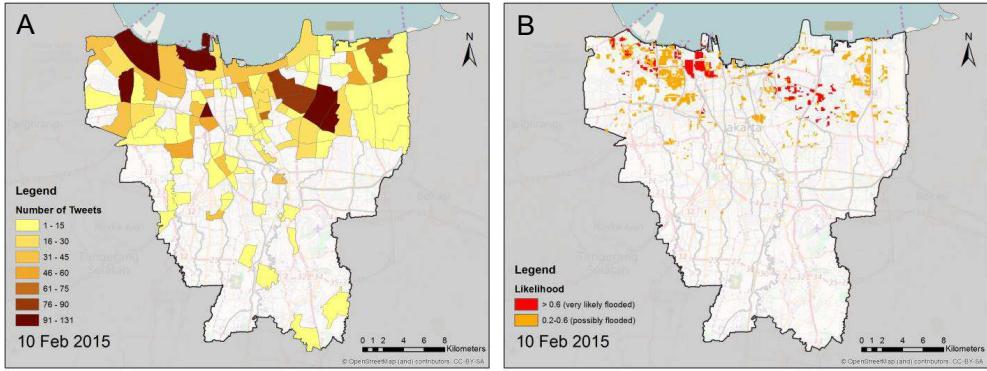


Figure 48: Mapping floods based on Twitter data. (A) Total number of flood-related tweets that contain a water depth. (B) Derived flood likelihood map. Figure from [851].

event. Then, they estimated the centers of events by applying location estimation methods such as particle filtering [843]. Their method can promptly detect 93% of earthquakes in Japan. Later, Sakaki et al. [844] developed a system based on the above method, which can send alarm e-mails about an earthquake to relevant people much faster than the official agency. Similarly, Earle et al. [845] developed an earthquake detection algorithm relying solely on tweets. They found that the peak of tweets containing “earthquake” is correlated with the event time. A short-term-average over long-term-average algorithm can effectively detect 48 globally-distributed earthquakes based on five months’ tweets. Their algorithm runs very fast, with 75% of detections accomplished within two minutes of the event time. For Australia and New Zealand, Robinson et al. [846] built a sensitive earthquake detector by monitoring earthquake-related tweets. They located earthquakes by examining tweets that contributed to an alert. Their detector can identify earthquake with an accuracy of about 0.81 in terms of F1 score at the best case. SM has multi-level functionalities during earthquakes such as interpersonal communications and information sharing [847], and thus it can be utilized to design effective monitoring and warning systems.

SM data have been increasingly used in monitoring and mapping floods in a timely manner. Vieweg et al. [848] analyzed tweets containing case-insensitive terms of the 2009 Red River Floods. They identified some high-level situational features of information generated during emergencies. Similarly, Cheong and Cheong [849] analyzed tweets generated during the Australian 2010-2011 floods and revealed interesting features of interactions between Twitter users during the crisis. They found that local authorities, officials and volunteers are influential players of the online communities. Later, de Albuquerque et al. [850] analyzed tweets generated during the 2013 River Elbe Flood. They found that the appearances of flood-related tweets show a general spatial pattern that flood-related tweets are strongly correlated with distance to flood events and relative water level. Many independent observations reporting the same flood on Twitter are more reliable. Following this principle, Eilander et al. [851] explored physical characteristics of floods and applied filtering and geo-statistical methods to assess the reliability of tweets over all flooded areas based on multiple observations. They developed an approach to construct a flood probability map. When tested in Jakarta, their approach can detect 93% of flood locations in the regions where people tweeted about water depth (see Figure 48). Recently, Arthur et al. [852] leveraged tweets to detect and locate flood events in the UK. They collected tweets containing flood-related terms and located flood events by analyzing many indicators such as mentioned place names and GPS coordinates. They produced high-quality flood event maps based on the relevant geotagged tweets and validated the flood maps by official data. Similarly, Li et al. [853] identified the spatio-temporal patterns of flood-related tweets for the 2015 South Carolina floods. They further proposed a kernel-based model to map the possibility of floods based on the water height mentioned in tweets.

In addition to tweets, data from other SM platforms can also be used to map and predict floods. Tkachenko et al. [854] analyzed tags of Flickr images uploaded during floods. They found that flood-related tags are correlated with hydrologically themed tags and then revealed connections between risk-signalling and generic environmental semantics. Further, they showed that environmental semantics derived from volunteered geographic data are helpful for improving flood warning. For the 2014 UK flood, Rosser et al. [855] collected geotagged photographs from Flickr

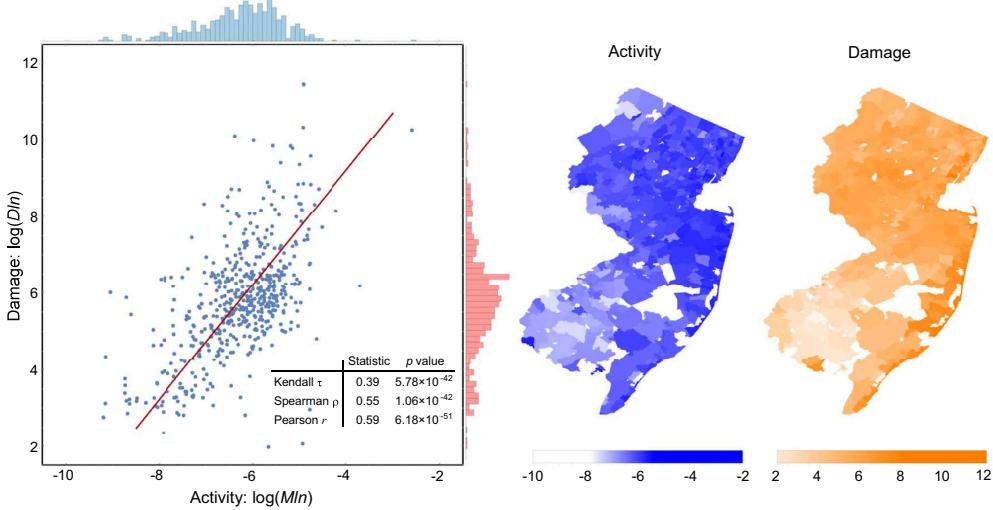


Figure 49: Correlations between Twitter activity and Hurricane Sandy damage and their spatial distributions. The per-capita Twitter activity and damage both follow a quasi log-normal distribution. The correlations between activity and damage is shown for New Jersey. Figure from [861].

and satellite imagery. They developed a Bayesian statistical model to estimate the probability of flood inundation by combining SM and remote sensing (RS) data. Their method can effectively predict spatial flood inundation with an AUC value over 0.93. Wang et al. [856] collected flood-related tweets from Twitter and flooding photos from the crowdsourcing platform MyCoast. They applied natural language processing (NLP) methods [857] to extract location names and flood depth from tweets and then employed convolutional neural networks (CNN) [20] to automatically classify flooding photos. They found that the combination of tweets and crowdsourcing photos can provide high-resolution flood monitoring. Recently, observations on many SM platforms have been used to map and model floods [858], showing the valuable means of rapidly estimated flood maps in improving situational awareness after floods.

The 2012 Hurricane Sandy disaster is one of the costliest disasters in the US history. Preis et al. [859] studied users' attention to Hurricane Sandy by examining photos on Flickr with related tags, titles or descriptions. They found that the moving average of the normalized number of photos related to Hurricane Sandy (under one day window) has a striking correlation (Kendall's  $\tau = -0.37$ ) [860] with the atmospheric pressure in New Jersey during the hurricane period, suggesting Flickr as a real-time sensor to track collective attention during emergencies. Based on 9.7 million geotagged tweets containing hurricane-related keywords, Kryvasheyev et al. [861] performed a multiscale analysis of Twitter activity before, during and after Hurricane Sandy. They found that the number of tweets containing keywords is strongly correlated with the peak of wind power on the day of hurricane landfall, and the magnitude of tweet activity increases with the proximity to the hurricane path. Moreover, there is a moderate correlation between the per-capita Twitter activity and the economic damage inflicted by Hurricane Sandy. The Kendall and Spearman rank correlations respectively reach 0.39 and 0.55 at the level of zip code tabulation areas (ZCTA) for New Jersey (see Figure 49). Moreover, they found that sentiment in Twitter is a weak predictor of economic damage. Gruebner et al. [862] sentimentally analyzed 344,957 geotagged tweets from the greater New York City (NYC) area and extracted basic emotions during Hurricane Sandy. They found that sadness is the most pronounced emotion during the whole hurricane period, while surprise and fear peak on the day of hurricane landfall. Further, they detected space-time clusters of excess risk of multiple basic emotions. Their work has taken a significant step towards improving mental health surveillance after disasters.

SM data have been applied to quantify human mobility perturbation and to improve resilience analysis of hurricane disasters. Wang and Taylor [863] studied the perturbation and resilience of human movement after Hurricane Sandy landed in NYC by analyzing tweets. They found that long-distance movements are significantly decreased during the first day. Moreover, human displacements follow a truncated power-law distribution with power-law exponent being 1.73 in the first day and being around 1.19 in the following 11 days. These results suggest that human mobility patterns are significantly perturbed by Hurricane Sandy. After further calculating the radius of gyration and the center

of mass of each individual's movement trajectory [55], they found a high correlation between values of these human mobility parameters under perturbation and those in steady states. These findings imply an inherent resilience of human mobility patterns during hurricanes and may be helpful for developing strategies to enable more effective evacuations. Based on content of geotagged tweets, Middleton et al. [864] developed a real-time crisis-mapping platform at the building- and street-level resolutions. After applied to Hurricane Sandy, their platform can produce the crisis mapping with an F1 value 0.53 compared to the official post-event impact assessment. Recently, Zou et al. [865] analyzed Twitter activities during Hurricane Sandy and showed an improved estimation on post-hurricane damage.

Mining real-time and large-scale SM data can also provide rapid situational awareness with precise locations for many other disasters, such as drought, wildfires and snowstorms. Tang et al. [866] found that governmental agencies used SM platforms as communication channels for information sharing during the 2014 California drought. In particular, Twitter plays an important role in expediting drought risk information dissemination. Boulton et al. [867] tested the feasibility of mapping wildfires from SM data. After analyzing wildfire-related posts collected from Twitter and Instagram, they found a positive correlation between SM activity and officially reported fire occurrences. Specifically, the correlations are  $r = 0.529$  for Twitter and  $r = 0.716$  for Instagram. Further, they analyzed the spatial and temporal features of the wildfire-related tweets and found that hotspots of wildfire-related Twitter activity are very likely to be the wildfire locations. Hong et al. [868] analyzed geotagged tweets generated during 18 snowstorms on the US east coast. They found that local governments used Twitter mainly for preparedness and response, while citizens used Twitter mainly for sharing their opinions about snowstorms. As a summary, SM is a rich data source that can be utilized to monitor disaster events, fasten damage assessment, improve disaster management and reduce economic losses.

## 6. Discussions

Ranging from international trade networks and global inequality of household income and individual wealth, from perception and prediction of socioeconomic status to prevention and protection of public health and natural disaster, through rich examples, this review shows how to dig out novel insights about socioeconomic development based on large-scale real data and advanced analysis tools like data mining and machine learning. These insights are usually not easy to be obtained by traditional methods.

At the same time, as an emerging branch of science, the studies of computational socioeconomics still confront some shortcomings and challenges. Methodologically speaking, the problems are twofold.

Firstly, the data quality, especially the authenticity of the data, cannot be fully guaranteed. In most world-known social media platforms, a considerable fraction of users are not human beings but artificial bots (named as *social bots* in literature) [869]. These bots do not just simply bring some noises, but they may be controlled by some commercial or academic institutions, and be assigned with some certain tasks. Therefore, these bots can largely affect the tendency of public sentiment and accelerate the spreading of rumors [870]. Accordingly, the results obtained from data sets polluted by social bots are probably far different from the reality. In addition, when a certain project has successfully drawn world-wide attention to its results (e.g., how to detect influenza epidemics using Google search query data [705]), some players in academic community may intentionally generate artificial data to disturb the reported models and algorithms (maybe not for manipulation, but for fun). Such smart jokes will also reduce the data authenticity and result in incorrect conclusions [709]. In a word, data are not collected from the pure land, and thus we'd better implement effective pretreatment to improve the quality of data before analysis and take into account the possible risks caused by noisy and unreal data before reporting any results to the public.

Secondly, the applicability and relevance of results are limited. First of all, through some data resources can cover a huge number of samples (even scaling with the whole population, such as mobile phones, Facebook, Twitter, WeChat, Weibo, etc.), these samples are not drawn randomly and thus cannot represent the whole population. For example, children, old and poor people are less engaged in the Internet and mobile Internet, resulting in the less chance to cover them by the above-mentioned data resources. Therefore, we have to carefully clarify the range of validity of the findings [871]. Secondly, socioeconomic problems are highly affected by local landscapes of religion, culture and politics, and thus a certain conclusion validated in one region may be not applicable in some other regions. For example, by analyzing the religion network consisted of believers in Weibo, Hu et al. [433] showed that religions in China are highly segregated, while about a half cross-religion links in Weibo are related to charitable issues. However, since the organization of religions in China is different from many other countries, whether the findings reported in

Ref. [433] are suitable for other countries asks for further validations. Lastly, even for the same dataset, the statistical regularities observed at the group level cannot be indiscriminately imitated at the individual level, or vice versa [142]. This is very similar to the Simpson's paradox [872] in statistics, where a trend appears in several different groups of data will disappear or reverse when these groups are combined.

In addition to the aforementioned shortcomings, we think the following five open issues are worth for future studies.

Firstly, we should try to design novel indices with a high ability in explanation and prediction. Currently, a considerable fraction of studies in computational socioeconomics applied new data resources and advanced analysis tools to estimate and predict routine socioeconomic indicators, such as GDP [100]. Such works are very valuable and easy to be accepted by the traditional socioeconomic community, while an obvious drawback is that any new methods cannot outperform the original statistical methods corresponding to the target indices. For example, even a complicated and advanced algorithm based on large-scale multiple-resource data can achieve an accuracy of 0.99 in estimating GDP, it is still worse than the original statistical method for GDP, whose accuracy is 1. Regarding this point, the designs of Economic Complexity Index [83] and Fitness [92] are successful examples. Very recently, satellite-based remote sensing data are also used to measure the level of manufacturing (see, for example, the China Satellite Manufacturing Index produced by SpaceKnow [873]). In a word, in addition to using novel data sources to estimate known data, we'd better design some new indices that make use of the novel datasets and well reflect valuable and important socioeconomic status behind the data. This issue still needs more attention and effort.

Secondly, it is valuable to re-evaluate the correctness and validation of traditional socioeconomic theories by some data-driven methods. Many traditional socioeconomic theories originate from simplified models of socioeconomic insights or extensions of known theories, which have not been validated by real data or have just been validated by small datasets in a highly limited range. It lacks comprehensive evaluation based on large-scale data that support in-depth analysis across different social formations, political systems, ideologies, culture traditions, economic levels, and so on. Using novel methods to critically evaluate traditional theories can also help computational socioeconomics to attract sufficient attention from traditional socioeconomics and thus to accelerate the integration of socioeconomics and computational socioeconomics.

Thirdly, we should reveal the underlying causal relationships and provide theoretical insights. The majority of findings reported in this review are only about the correlations between data features and targeted metrics. No matter how strong the correlation is and whether the data features are successfully utilized to predict the tendency of some targeted metrics, if one cannot find solid and robust causal relationships according to the inspiration from correlations, the theoretical value is limited [874]. Only if some future studies in computational socioeconomics could find out significant causal relationships via in-depth analysis of large-scale data, and then build up theoretical models that can withstand strict evaluations, the foundation of computational economics is considered to be solid.

Fourthly, we need to verify newly reported theories by controlled experiments. Controlled experiments play a central role in the methodology of socioeconomics and sociology. Although this review emphasizes on the studies based on unobtrusive data, it does not suggest the abandonment of controlled experiments. In fact, to design and implement controlled experiments is a very effective and sometimes necessary way towards uncovering solid causal relationships. Different from routine experiments, we can launch experiments covering huge population through the Internet and/or mobile Internet (e.g., Bond et al. [875] reported a 61-million-person experiment about social influence and political mobilization via Facebook). We can analyze large-scale data to find out potentially key factors, which can be utilized to better design and guide the following experiments.

Fifthly, we should try our best to find applications of theoretical and empirical analyses in practice. In despite of some effective methods, a couple of significant metrics and a few interesting conclusions developed by computational socioeconomics, overall speaking, the majority of known works have not found enough real applications or just generated some armchair strategies. However, the long-term value and vitality of computational socioeconomics depend on whether it can at least lead to some successes in practice. We suggest building up a system that can provide high-resolution and real-time monitoring about socioeconomic status and thus can detect possible risks at an early stage, so that governmental administrators, enterprisers, investors and other related people can make better decisions accordingly. This monitoring system should make use of multiple data sources, including the novel sources like mobile phones, satellites and social media platforms and the traditional sources like the data from economic censuses, questionnaire surveys and statistical yearbooks. Of particular importance, we suggest putting forward a series of data-driven policy suggestions (for example, the work by Alshamsi et al. [326] is a representative attempt but still far

from practical policies) and implementing these suggested policies in some qualified regions (even though the most critical issue is whether the heads of the corresponding governments support such blaze new trails), which is the way having the chance to make really big contributions. The classical economics failed to provide effective prescriptions to the world-wide poverty (for example, looking at the results of the huge efforts by the World Bank in fighting with the poverty of these so-called developing countries after the World War II [876, 877]), while if computational socioeconomic studies can find a new path to help impoverished people, its contribution to the world is tremendous.

We believe this review offers many valuable enlightenments to researchers in socioeconomics and other branches of social science. However, besides the booming development of computational socioeconomic, one should calmly notice that many important methods and conclusions introduced here still have not been accepted as a part of socioeconomic. Indeed, researchers doing works related to computational socioeconomic, currently being scattered in many disciplines, have not yet been seriously treated as challengers to the traditional socioeconomic methodology. In a word, we are still at the very early stage of the paradigm shifting of social science driven by big data and artificial intelligence. The way towards a quantitative version of social science is long and rugged, but undoubtedly, changes are happening and becoming increasingly fierce.

## Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (Grant Nos. 61433014, 61603074, 61673086 and 61703074) and the Science Promotion Programme of UESTC (No. Y03111023901014006). The authors acknowledge the support of the Swiss National Science Foundation (Grant No. 182498) during this collaboration. J.G. acknowledges the China Scholarship Council for a scholarship (No. 201606070051) and the Collective Learning Group at the MIT Media Lab for hosting.

## References

- [1] P. Ball, *Critical Mass: How One Thing Leads to Another*, Macmillan, London, UK, 2009.
- [2] A.-L. Barabási, *Bursts: The Hidden Patterns Behind Everything We Do, from Your E-mail to*, Penguin, New York, NY, USA, 2010.
- [3] R. L. Hughes, The flow of human crowds, *Annual Review of Fluid Mechanics* 35 (2003) 169–182.
- [4] D. Helbing, *Quantitative Sociodynamics: Stochastic Methods and Models of Social Interaction Processes*, 2nd Edition, Springer, Berlin, Heidelberg, 2010.
- [5] K. Popper, *The Logic of Scientific Discovery*, Routledge, New York, NY, USA, 2005.
- [6] N. Silver, *The Signal and the Noise: Why So Many Predictions Fail—but Some Don’t*, Penguin, New York, NY, USA, 2012.
- [7] N. N. Taleb, *The Black Swan: The Impact of the Highly Improbable*, Random House, New York, NY, USA, 2007.
- [8] D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, M. Van Alstyne, Computational social science, *Science* 323 (5915) (2009) 721–723.
- [9] D. V. Shah, J. N. Cappella, W. R. Neuman, Big data, digital media, and computational social science: Possibilities and perils, *ANNALS of the American Academy of Political and Social Science* 659 (1) (2015) 6–13.
- [10] R. J. Fisher, Social desirability bias and the validity of indirect questioning, *Journal of Consumer Research* 20 (2) (1993) 303–315.
- [11] V. Mayer-Schönberger, K. Cukier, *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, Houghton Mifflin Harcourt, New York, NY, USA, 2013.
- [12] P. F. Merenda, Toward a four-factor theory of temperament and/or personality, *Journal of Personality Assessment* 51 (3) (1987) 367–374.
- [13] S. D. Gosling, P. J. Rentfrow, W. B. Swann Jr, A very brief measure of the Big-Five personality domains, *Journal of Research in Personality* 37 (6) (2003) 504–528.
- [14] S. Battiston, J. D. Farmer, A. Flache, D. Garlaschelli, A. G. Haldane, H. Heesterbeek, C. Hommes, C. Jaeger, R. May, M. Scheffer, Complexity theory and financial regulation, *Science* 351 (6275) (2016) 818–819.
- [15] C. M. Reinhart, K. S. Rogoff, The aftermath of financial crises, *American Economic Review* 99 (2) (2009) 466–472.
- [16] J. Y. Lin, New structural economics: A framework for rethinking development, *World Bank Research Observer* 26 (2) (2011) 193–221.
- [17] M. S. Granovetter, The strength of weak ties, *American Journal of Sociology* 78 (6) (1973) 1360–1380.
- [18] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, A.-L. Barabási, Structure and tie strengths in mobile communication networks, *Proceedings of the National Academy of Sciences of the United States of America* 104 (18) (2007) 7332–7336.
- [19] P. S. Park, J. E. Blumenstock, M. W. Macy, The strength of long-range ties in population-scale social networks, *Science* 362 (6421) (2018) 1410–1413.
- [20] Y. Lecun, Y. Bengio, G. Hinton, Deep learning, *Nature* 521 (7553) (2015) 436–444.
- [21] J. E. Blumenstock, Fighting poverty with data, *Science* 353 (6301) (2016) 753–754.
- [22] S. Kuznets, Economic growth and income inequality, *American Economic Review* 45 (1) (1955) 1–28.
- [23] J. Gao, T. Zhou, Big data reveal the status of economic development, *Journal of University of Electronic Science and Technology of China* 45 (4) (2016) 625–633.
- [24] M. Ravallion, S. Chen, P. Sangraula, Dollar a day revisited, *World Bank Economic Review* 23 (2) (2009) 163–184.

- [25] M. Ravallion, How long will it take to lift one billion people out of poverty?, *World Bank Research Observer* 28 (2) (2013) 139–158.
- [26] D. Hulme, A. McKay, Identifying and measuring chronic poverty: Beyond monetary measures?, in: N. Kakwani, J. Silber (Eds.), *The Many Dimensions of Poverty*, Palgrave Macmillan, London, UK, 2007, pp. 187–214.
- [27] C. K. Paul, A. C. Mascarenhas, Remote sensing in development, *Science* 214 (4517) (1981) 139–145.
- [28] T. Ghosh, S. J. Anderson, C. D. Elvidge, P. C. Sutton, Using nighttime satellite imagery as a proxy measure of human well-being, *Sustainability* 5 (12) (2013) 4988–5019.
- [29] C. D. Elvidge, J. Safran, B. Tuttle, P. Sutton, P. Cinzano, D. Pettit, J. Arvesen, C. Small, Potential for global mapping of development via a nightsat mission, *GeoJournal* 69 (1–2) (2007) 45–53.
- [30] C. D. Elvidge, K. E. Baugh, E. A. Kihn, H. W. Kroehl, E. R. Davis, C. W. Davis, Relation between satellite observed visible-near infrared emissions, population, economic activity and electric power consumption, *International Journal of Remote Sensing* 18 (6) (1997) 1373–1379.
- [31] C. H. Doll, J.-P. Muller, C. D. Elvidge, Night-time imagery as a tool for global mapping of socioeconomic parameters and greenhouse gas emissions, *AMBIO: A Journal of the Human Environment* 29 (3) (2000) 157–162.
- [32] P. C. Sutton, R. Costanza, Global estimates of market and non-market values derived from nighttime satellite imagery, land cover, and ecosystem service valuation, *Ecological Economics* 41 (3) (2002) 509–527.
- [33] S. Ebener, C. Murray, A. Tandon, C. C. Elvidge, From wealth to health: Modelling the distribution of income per capita at the sub-national level using night-time light imagery, *International Journal of Health Geographics* 4 (1) (2005) 5.
- [34] A. M. Noor, V. A. Alegana, P. W. Gething, A. J. Tatem, R. W. Snow, Using remotely sensed night-time light as a proxy for poverty in Africa, *Population Health Metrics* 6 (2008) 5.
- [35] D. Rogers, T. Emwanu, T. Robinson, Poverty mapping in Uganda: An analysis using remotely sensed and other environmental data, PPLI Working Paper No. 36, Pro-Poor Livestock Policy Initiative, Rome, Italy (2006).
- [36] C. D. Elvidge, P. C. Sutton, T. Ghosh, B. T. Tuttle, K. E. Baugh, B. Bhaduri, E. Bright, A global poverty map derived from satellite data, *Computers and Geosciences* 35 (8) (2009) 1652–1660.
- [37] J. E. Dobson, E. A. Bright, P. R. Coleman, R. C. Durfee, B. A. Worley, Landscan: A global population database for estimating populations at risk, *Photogrammetric Engineering and Remote Sensing* 66 (7) (2000) 849–858.
- [38] T. Ghosh, R. L Powell, C. D Elvidge, K. E Baugh, P. C Sutton, S. Anderson, Shedding light on the global distribution of economic activity, *Open Geography Journal* 3 (1) (2010) 148–161.
- [39] J. V. Henderson, A. Storeygard, D. N. Weil, Measuring economic growth from outer space, *American Economic Review* 102 (2) (2012) 994–1028.
- [40] A. Mveyange, Night lights and regional income inequality in Africa, WIDER Working Paper Series No. 085, World Institute for Development Economic Research, Tokyo, Japan (2015).
- [41] S. Khandker, J. Haughton, *Handbook on Poverty and Inequality*, World Bank, Washington, D.C., USA, 2009.
- [42] P. Cauwels, N. Pestalozzi, D. Sornette, Dynamics and spatial distribution of global nighttime lights, *EPJ Data Science* 3 (2013) 2.
- [43] M. M. Bennett, L. C. Smith, Advances in using multitemporal night-time lights satellite imagery to detect, estimate, and monitor socioeconomic dynamics, *Remote Sensing of Environment* 192 (2017) 176–197.
- [44] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, S. Ermon, Combining satellite imagery and machine learning to predict poverty, *Science* 353 (6301) (2016) 790–794.
- [45] S. J. Pan, Q. Yang, A survey on transfer learning, *IEEE Transactions on Knowledge and Data Engineering* 22 (10) (2010) 1345–1359.
- [46] M. Xie, N. Jean, M. Burke, D. Lobell, S. Ermon, Transfer learning from deep features for remote sensing and poverty mapping, in: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI'16, AAAI Press, Palo Alto, CA, USA, 2016, pp. 3929–3935.
- [47] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, *Communications of the ACM* 60 (6) (2017) 84–90.
- [48] K. Chatfield, K. Simonyan, A. Vedaldi, A. Zisserman, Return of the devil in the details: Delving deep into convolutional nets, in: Proceedings of the British Machine Vision Conference, BMVA Press, Surrey, UK, 2014, pp. 1–12.
- [49] L. Sedda, A. J. Tatem, D. W. Morley, P. M. Atkinson, N. A. Wardrop, C. Pezzullo, A. Sorichetta, J. Kuleszo, D. J. Rogers, Poverty, health and satellite-derived vegetation indices: Their inter-spatial relationship in West Africa, *International Health* 7 (2) (2015) 99–106.
- [50] M. Imran, A. Stein, R. Zurita-Milla, Investigating rural poverty and marginality in Burkina Faso using remote sensing-based products, *International Journal of Applied Earth Observation and Geoinformation* 26 (2014) 322–334.
- [51] United Nations Global Pulse Lab Kampala. Measuring poverty with machine roof counting [online] (2019).  
URL <https://www.unglobalpulse.org/projects/measuring-poverty-machine-roof-counting>
- [52] G. R. Watmough, P. M. Atkinson, A. Saikia, C. W. Hutton, Understanding the evidence base for poverty–Environment relationships using remotely sensed satellite data: An example from Assam, India, *World Development* 78 (2016) 188–203.
- [53] W. Hong, X.-P. Han, T. Zhou, B.-H. Wang, Heavy-tailed statistics in short-message communication, *Chinese Physics Letters* 26 (2) (2009) 028902.
- [54] Z.-D. Zhao, H. Xia, M.-S. Shang, T. Zhou, Empirical analysis on the human dynamics of a large-scale short message communication system, *Chinese Physics Letters* 28 (6) (2011) 068901.
- [55] M. C. González, C. A. Hidalgo, A.-L. Barabási, Understanding individual human mobility patterns, *Nature* 453 (7196) (2008) 779–782.
- [56] J. Blumenstock, Y. Shen, N. Eagle, A method for estimating the relationship between phone use and wealth, in: QualMeetsQuant Workshop at the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development, Vol. 13, ACM Press, New York, NY, USA, 2010, pp. 114–125.
- [57] J. E. Blumenstock, N. Eagle, Divided we call: Disparities in access and use of mobile phones in Rwanda, *Information Technologies and International Development* 8 (2) (2012) 1–16.
- [58] J. E. Blumenstock, N. Eagle, M. Fafchamps, Airtime transfers and mobile communications: Evidence in the aftermath of natural disasters, *Journal of Development Economics* 120 (2016) 157–181.

- [59] T. Gutierrez, G. Krings, V. D. Blondel, Evaluating socio-economic state of a country analyzing airtime credit and mobile phone datasets, arXiv:1309.4496, 2013.
- [60] C. Smith, A. Mashhadi, L. Capra, Ubiquitous sensing for mapping poverty in developing countries, in: Proceedings of the Third Conference on the Analysis of Mobile Phone Datasets, NetMob 2013, MIT Media Lab, Cambridge, MA, USA, 2013, pp. 1–7.
- [61] United Nations Development Programme, Human Development Report 2010–20th Anniversary Edition. The real wealth of nations: pathways to human development, Tech. rep., United Nations Development Programme, New York, NY, USA (2010).
- [62] H. Mao, X. Shuai, Y.-Y. Ahn, J. Bollen, Quantifying socio-economic indicators in developing countries from mobile phone communication data: applications to Côte d'Ivoire, *EPJ Data Science* 4 (2015) 15.
- [63] S. Brin, L. Page, The anatomy of a large-scale hypertextual web search engine, *Computer Networks and ISDN Systems* 30 (1–7) (1998) 107–117.
- [64] S. Zhou, R. J. Mondragon, The rich-club phenomenon in the Internet topology, *IEEE Communications Letters* 8 (3) (2004) 180–182.
- [65] A. Flammini, V. Colizza, M. Serrano, A. Vespignani, Rich-club ordering in complex networks, *Nature Physics* 2 (3) (2006) 110–115.
- [66] J. Blumenstock, G. Cadamuro, R. On, Predicting poverty and wealth from mobile phone metadata, *Science* 350 (6264) (2015) 1073–1076.
- [67] H. Zou, T. Hastie, Regularization and variable selection via the Elastic Net, *Journal of the Royal Statistical Society. Series B (Methodological)* 67 (2) (2005) 301–320.
- [68] J. E. Steele, P. R. Sundsøy, C. Pezzulo, V. A. Alegana, T. J. Bird, J. Blumenstock, J. Bjelland, K. Engø-Monsen, Y.-A. de Montjoye, A. M. Iqbal, K. N. Hadiuzzaman, Mapping poverty using mobile phone and satellite data, *Journal of the Royal Society Interface* 14 (127) (2017) 20160690.
- [69] M. Blangiardo, M. Cameletti, Spatial and Spatio-temporal Bayesian Models with R-INLA, John Wiley & Sons, New York, NY, USA, 2015.
- [70] A. Okabe, B. Boots, K. Sugihara, Spatial Tessellations: Concepts and Applications of Voronoi Diagrams, John Wiley & Sons, New York, NY, USA, 1992.
- [71] C. Njuguna, P. McSharry, Constructing spatiotemporal poverty indices from big data, *Journal of Business Research* 70 (2017) 318–327.
- [72] United Nations Global Pulse, Building proxy indicators of national wellbeing with postal data, Global Pulse Project Series No. 22, United Nations Global Pulse, New York, NY, USA (2016).
- [73] D. Hristova, A. Rutherford, J. Anson, M. Luengo-Oroz, C. Mascolo, The international postal network and other global flows as proxies for national wellbeing, *PLoS ONE* 11 (6) (2016) e0155976.
- [74] M. Kivelä, A. Arenas, M. Barthélémy, J. P. Gleeson, Y. Moreno, M. A. Porter, Multilayer networks, *Journal of Complex Networks* 2 (3) (2014) 203–271.
- [75] F. Battiston, V. Nicosia, V. Latora, Structural measures for multiplex networks, *Physical Review E* 89 (3) (2014) 032804.
- [76] R. Hausmann, J. Hwang, D. Rodrik, What you export matters, *Journal of Economic Growth* 12 (1) (2007) 1–25.
- [77] C. A. Hidalgo, B. Klenger, A.-L. Barabási, R. Hausmann, The product space conditions the development of nations, *Science* 317 (5837) (2007) 482–487.
- [78] B. Balassa, Trade liberalisation and “revealed” comparative advantage, *Manchester School* 33 (2) (1965) 99–123.
- [79] S. P. Borgatti, M. G. Everett, Models of core/periphery structures, *Social Networks* 21 (4) (2000) 375–395.
- [80] P. Holme, Core-periphery organization of complex networks, *Physical Review E* 72 (4) (2005) 046111.
- [81] F. Neffke, M. Henning, R. Boschma, How do regions diversify over time? Industry relatedness and the development of new growth paths in regions, *Economic Geography* 87 (3) (2011) 237–265.
- [82] A. Abdon, J. Felipe, The product space: What does it say about the opportunities for growth and structural transformation of Sub-Saharan Africa?, Working Papers No. 670, Levy Economics Institute, Annandale-on-Hudson, NY, USA (2011).
- [83] C. A. Hidalgo, R. Hausmann, The building blocks of economic complexity, *Proceedings of the National Academy of Sciences of the United States of America* 106 (26) (2009) 10570–10575.
- [84] R. Hausmann, C. A. Hidalgo, S. Bustos, M. Coscia, A. Simoes, M. A. Yildirim, The Atlas of Economic Complexity: Mapping Paths to Prosperity, MIT Press, Cambridge, MA, USA, 2014.
- [85] R. Hausmann, C. A. Hidalgo, The network structure of economic output, *Journal of Economic Growth* 16 (4) (2011) 309–342.
- [86] S. Bustos, C. Gomez, R. Hausmann, C. A. Hidalgo, The dynamics of nestedness predicts the evolution of industrial ecosystems, *PLoS ONE* 7 (11) (2012) e49393.
- [87] B. D. Patterson, W. Atmar, Nested subsets and the structure of insular mammalian faunas and archipelagos, *Biological Journal of the Linnean Society* 28 (1–2) (1986) 65–82.
- [88] J. Bascompte, P. Jordano, C. J. Melián, J. M. Olesen, The nested assembly of plantanimal mutualistic networks, *Proceedings of the National Academy of Sciences of the United States of America* 100 (16) (2003) 9383–9387.
- [89] J. H. Lin, C. Tessone, M. Mariani, Nestedness maximization in complex networks through the fitness-complexity algorithm, *Entropy* 20 (10) (2018) 768.
- [90] J. Felipe, U. Kumar, A. Abdon, M. Bacate, Product complexity and economic development, *Structural Change and Economic Dynamics* 23 (1) (2012) 36–68.
- [91] G. Caldarelli, M. Cristelli, A. Gabrielli, L. Pietronero, A. Scala, A. Tacchella, A network analysis of countries’ export flows: Firm grounds for the building blocks of the economy, *PLoS ONE* 7 (10) (2012) e47278.
- [92] A. Tacchella, M. Cristelli, G. Caldarelli, A. Gabrielli, L. Pietronero, A new metrics for countries’ fitness and products’ complexity, *Scientific Reports* 2 (2012) 723.
- [93] J. M. Kleinberg, Authoritative sources in a hyperlinked environment, *Journal of the ACM* 46 (5) (1999) 604–632.
- [94] Y.-B. Zhou, T. Lei, T. Zhou, A robust ranking algorithm to spamming, *EPL (Europhysics Letters)* 94 (4) (2011) 48002.
- [95] E. Pugliese, A. Zaccaria, L. Pietronero, On the convergence of the Fitness-Complexity algorithm, *European Physical Journal Special Topics* 225 (10) (2016) 1893–1911.
- [96] M. Cristelli, A. Gabrielli, A. Tacchella, G. Caldarelli, L. Pietronero, Measuring the intangibles: A metrics for the economic complexity of countries and products, *PLoS ONE* 8 (8) (2013) e70726.
- [97] A. Tacchella, M. Cristelli, G. Caldarelli, A. Gabrielli, L. Pietronero, Economic complexity: Conceptual grounding of a new metrics for

- global competitiveness, *Journal of Economic Dynamics and Control* 37 (8) (2013) 1683–1691.
- [98] M. Cristelli, A. Tacchella, L. Pietronero, The heterogeneous dynamics of economic complexity, *PLoS ONE* 10 (2) (2015) e0117174.
- [99] E. N. Lorenz, Atmospheric predictability as revealed by naturally occurring analogues, *Journal of the Atmospheric Sciences* 26 (4) (1969) 636–646.
- [100] A. Tacchella, D. Mazzilli, L. Pietronero, A dynamical systems approach to gross domestic product forecasting, *Nature Physics* 14 (8) (2018) 861–865.
- [101] International Monetary Fund, World economic outlook: October 2016, subdued demand, symptoms and remedies, Tech. rep., International Monetary Fund, Washington, D.C., USA (2016).
- [102] M. S. Mariani, A. Vidmer, M. Medo, Y.-C. Zhang, Measuring economic complexity of countries and products: Which metric to use?, *European Physical Journal B* 88 (11) (2015) 293.
- [103] R.-J. Wu, G.-Y. Shi, Y.-C. Zhang, M. S. Mariani, The mathematics of non-linear metrics for nested networks, *Physica A: Statistical Mechanics and its Applications* 460 (2016) 254–269.
- [104] G. Morrison, S. V. Buldyrev, M. Imbruno, O. A. Doria Arrieta, A. Rungi, M. Riccaboni, F. Pammolli, On economic complexity and the fitness of nations, *Scientific Reports* 7 (2017) 15332.
- [105] A.-L. Barabási, R. Albert, Emergence of scaling in random networks, *Science* 286 (5439) (1999) 509–512.
- [106] R. Foschi, M. Riccaboni, S. Schiavo, Preferential attachment in multiple trade networks, *Physical Review E* 90 (2) (2014) 022817.
- [107] A. Zaccaria, M. Cristelli, A. Tacchella, L. Pietronero, How the taxonomy of products drives the economic development of countries, *PLoS ONE* 9 (12) (2014) e113770.
- [108] T. Zhou, J. Ren, M. Medo, Y.-C. Zhang, Bipartite network projection and personal recommendation, *Physical Review E* 76 (4) (2007) 046115.
- [109] F. Saracco, R. D. Clemente, A. Gabrielli, L. Pietronero, From innovation to diversification: A simple competitive model, *PLoS ONE* 10 (11) (2015) e0140420.
- [110] A. Zaccaria, M. Cristelli, R. Kupers, A. Tacchella, L. Pietronero, A case study for a new metrics for economic complexity: The Netherlands, *Journal of Economic Interaction and Coordination* 11 (1) (2016) 151–169.
- [111] V. Stojkoski, Z. Utkovski, L. Kocarev, The impact of services on economic complexity: Service sophistication as route for economic growth, *PLoS ONE* 11 (8) (2016) e0161633.
- [112] D. Hartmann, M. R. Guevara, C. Jara-Figueroa, M. Aristarán, C. A. Hidalgo, Linking economic complexity, institutions and income inequality, *World Development* 93 (2017) 75–93.
- [113] P. Mealy, J. D. Farmer, A. Teytelboym, Interpreting economic complexity, *Science Advances* 5 (1) (2019) eaau1705.
- [114] E. Pugliese, G. L. Chiarotti, A. Zaccaria, L. Pietronero, Complex economies have a lateral escape from the poverty trap, *PLoS ONE* 12 (1) (2017) e0168540.
- [115] A. Sbardella, E. Pugliese, L. Pietronero, Economic development and wage inequality: A complex system analysis, *PLoS ONE* 12 (9) (2017) e0182774.
- [116] A. J. G. Simoes, C. A. Hidalgo, The economic complexity observatory: An analytical tool for understanding the dynamics of economic development, in: *Proceedings of the 17th AAAI Conference on Scalable Integration of Analytics and Visualization, AAAIWS'11-17*, AAAI Press, Palo Alto, CA, USA, 2011, pp. 39–42.
- [117] C. Linard, A. J. Tatem, Large-scale spatial population databases in infectious disease research, *International Journal of Health Geographics* 11 (1) (2012) 7.
- [118] N. N. Patel, F. R. Stevens, Z. Huang, A. E. Gaughan, I. Elyazar, A. J. Tatem, Improving large area population mapping using geotweet densities, *Transactions in GIS* 21 (2) (2017) 317–331.
- [119] W. Tobler, U. Deichmann, J. Gottsegen, K. Maloy, World population in a grid of spherical quadrilaterals, *International Journal of Population Geography* 3 (3) (1997) 203–225.
- [120] B. Bhaduri, E. Bright, P. Coleman, M. L. Urban, LandScan USA: A high-resolution geospatial and temporal modeling approach for population distribution and dynamics, *Geojournal* 69 (1–2) (2007) 103–117.
- [121] A. J. Tatem, A. M. Noor, C. von Hagen, A. Di Gregorio, S. I. Hay, High resolution population maps for low income nations: Combining land cover and census in East Africa, *PLoS ONE* 2 (12) (2007) e1298.
- [122] A. Cheriyadat, E. Bright, Mapping of settlements in high-resolution satellite imagery using high performance computing, *Geojournal* 69 (1–2) (2007) 119–129.
- [123] M. de Martinao, F. Causa, S. B. Serpico, Classification of optical high resolution images in urban environment using spectral and textural information, in: *2003 IEEE International Geoscience and Remote Sensing Symposium*, Vol. 1, 2003, pp. 467–469.
- [124] Y. Liao, J. Wang, B. Meng, X. Li, Integration of GP and GA for mapping population distribution, *International Journal of Geographical Information Science* 24 (1) (2010) 47–67.
- [125] J. K. Kishore, L. M. Patnaik, V. Mani, V. K. Agrawal, Genetic programming based pattern classification with feature space partitioning, *Information Sciences* 131 (1–4) (2001) 65–86.
- [126] J. H. Holland, *Adaptation in Natural and Artificial Systems: An introductory analysis with applications to biology, control, and artificial intelligence*, MIT Press, Cambridge, MA, USA, 1992.
- [127] C. Deng, C. Wu, L. Wang, Improving the housing-unit method for small-area population estimation using remote-sensing and GIS information, *International Journal of Remote Sensing* 31 (21) (2010) 5673–5688.
- [128] A. E. Gaughan, F. R. Stevens, C. Linard, P. Jia, A. J. Tatem, High resolution population distribution maps for Southeast Asia in 2010 and 2015, *PLoS ONE* 8 (2) (2013) e55882.
- [129] F. R. Stevens, A. E. Gaughan, C. Linard, A. J. Tatem, Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data, *PLoS ONE* 10 (2) (2015) e0107042.
- [130] N. N. Patel, E. Angiuli, P. Gamba, A. Gaughan, G. Lisini, F. R. Stevens, A. J. Tatem, G. Trianni, Multitemporal settlement and population mapping from Landsat using Google Earth Engine, *International Journal of Applied Earth Observation and Geoinformation* 35 (2015) 199–208.

- [131] R. M. Pulselli, P. Romano, C. Ratti, E. Tiezzi, Computing urban mobile landscapes through monitoring population density based on cell-phone chatting, *International Journal of Design & Nature and Ecodynamics* 3 (2) (2008) 121–134.
- [132] Y.-F. Dan, Z.-S. He, A dynamic model for urban population density estimation using mobile phone location data, in: 2010 the 5th IEEE Conference on Industrial Electronics and Applications, ICIEA'10, IEEE Press, 2010, pp. 1429–1433.
- [133] S. Lloyd, Least squares quantization in pcm, *IEEE Transactions on Information Theory* 28 (2) (1982) 129–137.
- [134] C. Kang, Y. Liu, X. Ma, L. Wu, Towards estimating urban population distributions from mobile call data, *Journal of Urban Technology* 19 (4) (2012) 3–21.
- [135] P. Deville, C. Linard, S. Martin, M. Gilbert, F. R. Stevens, A. E. Gaughan, V. D. Blondel, A. J. Tatem, Dynamic population mapping using mobile phone data, *Proceedings of the National Academy of Sciences of the United States of America* 111 (45) (2014) 15888–15893.
- [136] R. W. Douglass, D. A. Meyer, M. Ram, D. Rideout, D. Song, High resolution population estimates from telecommunications data, *EPJ Data Science* 4 (2015) 4.
- [137] A. Lulli, L. Gabrielli, P. Dazzi, M. Dell'Amico, P. Michiardi, M. Nanni, L. Ricci, Improving population estimation from mobile calls: A clustering approach, in: 2016 IEEE Symposium on Computers and Communication, ISCC'16, IEEE Press, 2016, pp. 1097–1102.
- [138] G. Khodabandeh, V. Gauthier, M. El-Yacoubi, M. Fiore, Population estimation from mobile network traffic metadata, in: 2016 IEEE 17th International Symposium on A World of Wireless, Mobile and Multimedia Networks, WoWMoM'16, IEEE Press, 2016, pp. 1–9.
- [139] M. Kosinski, D. Stillwell, T. Graepel, Private traits and attributes are predictable from digital records of human behavior, *Proceedings of the National Academy of Sciences of the United States of America* 110 (15) (2013) 5802–2805.
- [140] M. Tsavli, P. S. Efraimidis, V. Katos, L. Mitrou, Reengineering the user: Privacy concerns about personal data on smartphones, *Information and Computer Security* 23 (4) (2015) 394–405.
- [141] K. Leetaru, S. Wang, G. Cao, A. Padmanabhan, E. Shook, Mapping the global Twitter heartbeat: The geography of Twitter, *First Monday* 18 (5) (2013) 4366.
- [142] Z. Yang, D. Lian, N. J. Yuan, X. Xie, Y. Rui, T. Zhou, Indigenization of urban mobility, *Physica A: Statistical Mechanics and its Applications* 469 (2017) 232–243.
- [143] Y. Yao, X. Liu, X. Li, J. Zhang, Z. Liang, K. Mai, Y. Zhang, Mapping fine-scale population distributions at the building level by integrating multisource geospatial big data, *International Journal of Geographical Information Science* 31 (6) (2017) 1220–1244.
- [144] L. Breiman, Random forests, *Machine Learning* 45 (1) (2001) 5–32.
- [145] A. J. Tatem, WorldPop, open data for spatial demography, *Scientific Data* 4 (2017) 170004.
- [146] A. Sorichetta, G. M. Hornby, F. R. Stevens, A. E. Gaughan, C. Linard, A. J. Tatem, High-resolution gridded population datasets for Latin America and the Caribbean in 2010, 2015, and 2020, *Scientific Data* 2 (2015) 150045.
- [147] A. E. Gaughan, F. R. Stevens, Z. Huang, J. J. Nieves, A. Sorichetta, S. Lai, X. Ye, C. Linard, G. M. Hornby, S. I. Hay, Spatiotemporal patterns of population in mainland China, 1990 to 2010, *Scientific Data* 3 (2016) 160005.
- [148] C. T. Lloyd, A. Sorichetta, A. J. Tatem, High resolution global gridded data for use in population studies, *Scientific Data* 4 (2017) 170001.
- [149] M. Boyd, Family and personal networks in international migration: Recent developments and new agendas, *International Migration Review* 23 (3) (1989) 638–670.
- [150] R. M. Friedberg, J. Hunt, The impact of immigrants on host country wages, employment and growth, *Journal of Economic Perspectives* 9 (2) (1995) 23–44.
- [151] J. de Beer, J. Raymer, R. van der Erf, L. van Wissen, Overcoming the problems of inconsistent international migration data: A new method applied to flows in Europe, *European Journal of Population* 26 (4) (2010) 459–481.
- [152] J. Raymer, G. Abel, P. W. F. Smith, Combining census and registration data to estimate detailed elderly migration flows in England and Wales, *Journal of the Royal Statistical Society: Series A* 170 (4) (2007) 891–908.
- [153] F. Willekens, Modeling approaches to the indirect estimation of migration flows: From entropy to EM, *Mathematical Population Studies* 7 (3) (1999) 239–278.
- [154] J. E. Cohen, M. Roig, D. C. Reuman, C. GoGwilt, International migration beyond gravity: A statistical model for use in population projections, *Proceedings of the National Academy of Sciences of the United States of America* 105 (40) (2008) 15269–15274.
- [155] P. McCullagh, Generalized linear models, *European Journal of Operational Research* 16 (3) (1989) 285–292.
- [156] G. J. Abel, N. Sander, Quantifying global international migration flows, *Science* 343 (6178) (2014) 1520–1522.
- [157] W. E. Deming, F. F. Stephan, On a least squares adjustment of a sampled frequency table when the expected marginal totals are known, *Annals of Mathematical Statistics* 11 (4) (1940) 427–444.
- [158] E. Zagheni, I. Weber, You are where you e-mail: Using e-mail data to estimate international migration rates, in: Proceedings of the 4th Annual ACM Web Science Conference, WebSci'12, ACM Press, New York, NY, USA, 2012, pp. 348–351.
- [159] B. State, I. Weber, E. Zagheni, Studying inter-national mobility through IP geolocation, in: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, WSDM'13, ACM Press, New York, NY, USA, 2013, pp. 265–274.
- [160] B. State, M. Rodriguez, D. Helbing, E. Zagheni, Migration of professionals to the U.S., in: Proceedings of the 6th International Conference on Social Informatics, SocInfo 2014, Springer, Cham, Switzerland, 2014, pp. 531–543.
- [161] R. Kikas, M. Dumas, A. Saabas, Explaining international migration in the Skype network: The role of social network features, in: Proceedings of the 1st ACM Workshop on Social Media World Sensors, SiDEWays'15, ACM Press, New York, NY, USA, 2015, pp. 17–22.
- [162] D. Barchiesi, H. S. Moat, C. Alis, S. Bishop, T. Preis, Quantifying international travel flows using Flickr, *PLoS ONE* 10 (7) (2015) e0128470.
- [163] B. Hawelka, I. Sitko, E. Beinat, S. Sobolevsky, P. Kazakopoulos, C. Ratti, Geo-located Twitter as proxy for global mobility patterns, *Cartography and Geographic Information Science* 41 (3) (2014) 260–271.
- [164] E. Zagheni, V. R. K. Garimella, I. Weber, B. State, Inferring international and internal migration patterns from Twitter data, in: Proceedings of the 23rd International Conference on World Wide Web, WWW'14 Companion, ACM Press, New York, NY, USA, 2014, pp. 439–444.
- [165] M. Bertrand, E. Duflo, S. Mullainathan, How much should we trust differences-in-differences estimates?, *Quarterly Journal of Economics* 119 (1) (2004) 249–275.
- [166] G. Fagiolo, M. Mastrorillo, International migration network: Topology and modeling, *Physical Review E* 88 (1) (2013) 012812.
- [167] G. Fagiolo, M. Mastrorillo, Does human migration affect international trade? A complex-network perspective, *PLoS ONE* 9 (5) (2014)

e97331.

- [168] J. Lee, J. Carling, P. Orrenius, The International Migration Review at 50: Reflecting on half a century of international migration research and looking ahead, *International Migration Review* 48 (2014) S3–S36.
- [169] Global Migration Group, Handbook for improving the production and use of migration data for development, Tech. rep., Global Migration Group, World Bank, Washington, D.C., USA (2016).
- [170] C. A. Bail, The cultural environment: Measuring culture with big data, *Theory and Society* 43 (3–4) (2014) 465–482.
- [171] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, E. L. Aiden, Quantitative analysis of culture using millions of digitized books, *Science* 331 (6014) (2011) 176–182.
- [172] J. Berger, E. T. Bradlow, A. Braunstein, Y. Zhang, From Karen to Katie: Using baby names to understand cultural evolution, *Psychological Science* 23 (10) (2012) 1067.
- [173] S. Ronen, B. Gonçalves, K. Z. Hu, A. Vespignani, S. Pinker, C. A. Hidalgo, Links that speak: The global language network and its association with global fame, *Proceedings of the National Academy of Sciences of the United States of America* 111 (52) (2014) E5616–E5622.
- [174] Y.-X. Zhu, J. Huang, Z.-K. Zhang, Q.-M. Zhang, T. Zhou, Y.-Y. Ahn, Geography and similarity of regional cuisines in China, *PLoS ONE* 8 (11) (2013) e79161.
- [175] A. Z. Yu, S. Ronen, K. Hu, T. Lu, C. A. Hidalgo, Pantheon 1.0, a manually verified dataset of globally famous biographies, *Scientific Data* 3 (2016) 150075.
- [176] J. Eisenstein, B. O'Connor, N. A. Smith, E. P. Xing, Diffusion of lexical change in social media, *PLoS ONE* 9 (11) (2014) e113114.
- [177] R. K. Larson, V. Déprez, H. Yamakido, *The Evolution of Human Language: Biolinguistic Perspectives*, Cambridge University Press, New York, NY, USA, 2010.
- [178] R. Zeng, P. M. Greenfield, Cultural evolution over the last 40 years in China: Using the Google Ngram Viewer to study implications of social and political change for cultural values, *International Journal of Psychology* 50 (1) (2015) 47–55.
- [179] B. Yucesoy, X. Wang, J. Huang, A.-L. Barabási, Success in books: A big data approach to bestsellers, *EPJ Data Science* 7 (2018) 7.
- [180] M. Schich, C. Song, Y.-Y. Ahn, A. Mirsky, M. Martino, A.-L. Barabási, D. Helbing, A network framework of cultural history, *Science* 345 (6196) (2014) 558–562.
- [181] D. Yang, D. Zhang, B. Qu, Participatory cultural mapping based on collective behavior data in location-based social networks, *ACM Transactions on Intelligent Systems and Technology* 7 (3) (2016) 30.
- [182] M. W. Hahn, R. A. Bentley, Drift as a mechanism for cultural change: An example from baby names, *Proceedings of the Royal Society of London B: Biological Sciences* 270 (Suppl 1) (2003) S120–S123.
- [183] R. A. Bentley, C. P. Lipo, H. A. Herzog, M. W. Hahn, Regular rates of popular culture change reflect random copying, *Evolution and Human Behavior* 28 (3) (2007) 151–158.
- [184] N. Xi, Z.-K. Zhang, Y.-C. Zhang, Z. Ge, L. She, K. Zhang, Cultural evolution: The case of babies' first names, *Physica A: Statistical Mechanics and its Applications* 406 (2014) 139–144.
- [185] P. Barucca, J. Rocchi, E. Marinari, G. Parisi, F. Riccitetenghi, Cross-correlations of American baby names, *Proceedings of the National Academy of Sciences of the United States of America* 112 (26) (2015) 7943–7947.
- [186] B. J. Kim, S. M. Park, Distribution of Korean family names, *Physica A: Statistical Mechanics and its Applications* 347 (2005) 683–694.
- [187] M. J. Lee, W. S. Jo, I. G. Yi, S. K. Baek, B. J. Kim, Evolution of popularity in given names, *Physica A: Statistical Mechanics and its Applications* 443 (2016) 415–422.
- [188] M. A. Nowak, N. L. Komarova, P. Niyogi, Computational and evolutionary aspects of language, *Nature* 417 (6889) (2002) 611–617.
- [189] D. M. Abrams, S. H. Strogatz, Linguistics: Modelling the dynamics of language death, *Nature* 424 (6951) (2003) 900–900.
- [190] E. Lieberman, J. B. Michel, J. Jackson, T. Tang, M. A. Nowak, Quantifying the evolutionary dynamics of language, *Nature* 449 (7163) (2007) 713–716.
- [191] M. G. Newberry, C. A. Ahern, R. Clark, J. B. Plotkin, Detecting evolutionary forces in language change, *Nature* 551 (7679) (2017) 223–226.
- [192] G. K. Zipf, Human behavior and the principle of least effort, *American Journal of Sociology* 110 (110) (1949) 306–306.
- [193] H. S. Heaps, *Information Retrieval: Computational and Theoretical Aspects*, Academic Press, Orlando, FL, USA, 1978.
- [194] Z.-K. Zhang, L. Lü, J.-G. Liu, T. Zhou, Empirical analysis on a keyword-based semantic system, *European Physical Journal B* 66 (4) (2008) 557–561.
- [195] L. Lü, Z.-K. Zhang, T. Zhou, Zipf's law leads to Heaps' law: Analyzing their relation in finite-size systems, *PLoS ONE* 5 (12) (2010) e14139.
- [196] A. M. Petersen, J. N. Tenenbaum, S. Havlin, H. E. Stanley, M. Perc, Languages cool as they expand: Allometric scaling and the decreasing need for new words, *Scientific Reports* 2 (2012) 943.
- [197] M. Gerlach, E. G. Altmann, Stochastic model for the vocabulary growth in natural languages, *Physical Review X* 3 (2) (2013) 021006.
- [198] E. A. Pechenick, C. M. Danforth, P. S. Dodds, Is language evolution grinding to a halt? The scaling of lexical turbulence in English fiction suggests it is not, *Journal of Computational Science* 21 (2017) 24–37.
- [199] L. Lü, Z.-K. Zhang, T. Zhou, Deviation of Zipf's and Heaps' Laws in human languages with limited dictionary sizes, *Scientific Reports* 3 (2013) 1082.
- [200] W. Deng, A. E. Allahverdyan, B. Li, Q. A. Wang, Rank-frequency relation for Chinese characters, *European Physical Journal B* 87 (2) (2014) 47.
- [201] X.-Y. Yan, P. Minnhagen, Maximum entropy, word-frequency, Chinese characters, and multiple meanings, *PLoS ONE* 10 (5) (2015) e0125592.
- [202] X. Yan, Y. Fan, Z. Di, S. Havlin, J. Wu, Efficient learning strategy of Chinese characters based on network approach, *PLoS ONE* 8 (8) (2013) e69745.
- [203] J. Bryden, S. Funk, V. A. Jansen, Word usage mirrors community structure in the online social network Twitter, *EPJ Data Science* 2 (2013) 3.
- [204] A. Samoilenco, F. Karimi, D. Edler, J. Kunegis, M. Strohmaier, Linguistic neighbourhoods: Explaining cultural borders on Wikipedia through multilingual co-editing activity, *EPJ Data Science* 5 (2016) 9.

- [205] D. Mocanu, A. Baronchelli, N. Perra, B. Gonçalves, Q. Zhang, A. Vespignani, The Twitter of babel: Mapping world languages through microblogging platforms, *PLoS ONE* 8 (4) (2013) e61981.
- [206] B. Gonçalves, L. Loureiro-Porto, J. J. Ramasco, D. Sánchez, Mapping the Americanization of English in space and time, *PLoS ONE* 13 (5) (2018) e0197741.
- [207] C. Counihan, P. van Esterik, *Food and Culture: A Reader*, 3rd Edition, Routledge, New York, NY, USA, 2013.
- [208] Y.-Y. Ahn, S. E. Ahnert, J. P. Bagrow, A.-L. Barabási, Flavor network and the principles of food pairing, *Scientific Reports* 1 (7377) (2011) 196.
- [209] H. Blumenthal, *The Big Fat Duck Cookbook*, Bloomsbury, London, UK, 2008.
- [210] C. Wagner, P. Singer, M. Strohmaier, The nature and evolution of online food preferences, *EPJ Data Science* 3 (2014) 38.
- [211] S. Abbar, Y. Mejova, I. Weber, You tweet what you eat: Studying food consumption through Twitter, in: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI'15, ACM, New York, NY, USA, 2015, pp. 3197–3206.
- [212] P. Laufer, C. Wagner, F. Flöck, M. Strohmaier, Mining cross-cultural relations from Wikipedia: A study of 31 European food cultures, in: Proceedings of the ACM Web Science Conference, WebSci'15, ACM, New York, NY, USA, 2015, p. 3.
- [213] P. C. Sutton, C. D. Elvidge, T. Ghosh, Estimation of gross domestic product at sub-national scales using nighttime satellite imagery, *International Journal of Ecological Economics and Statistics* 8 (S07) (2007) 5–21.
- [214] C. D. Elvidge, M. L. Imhoff, K. E. Baugh, V. R. Hobson, I. Nelson, J. Safran, J. B. Dietz, B. T. Tuttle, Night-time lights of the world: 1994–1995, *ISPRS Journal of Photogrammetry and Remote Sensing* 56 (2) (2001) 81–99.
- [215] T. K. Chand, K. Badarinath, C. Elvidge, B. Tuttle, Spatial characterization of electrical power consumption patterns over India using temporal DMSP-OLS night-time satellite data, *International Journal of Remote Sensing* 30 (3) (2009) 647–661.
- [216] P. Propastin, M. Kappas, Assessing satellite-observed nighttime lights for monitoring socioeconomic parameters in the Republic of Kazakhstan, *GIScience and Remote Sensing* 49 (4) (2012) 538–557.
- [217] X. Chen, W. D. Nordhaus, Using luminosity data as a proxy for economic statistics, *Proceedings of the National Academy of Sciences of the United States of America* 108 (21) (2011) 8589–8594.
- [218] T. Ma, C. Zhou, T. Pei, S. Haynie, J. Fan, Quantitative estimation of urbanization dynamics using time series of DMSP/OLS nighttime light data: A comparative case study from China's cities, *Remote Sensing of Environment* 124 (2012) 99–107.
- [219] C. Mellander, J. Lobo, K. Stolarick, Z. Matheson, Night-time light data: A good proxy measure for economic activity?, *PLoS ONE* 10 (10) (2015) e0139779.
- [220] N. Zhao, N. Currit, E. Samson, Net primary production and gross domestic product in China derived from satellite imagery, *Ecological Economics* 70 (5) (2011) 921–928.
- [221] M. Zhao, W. Cheng, C. Zhou, M. Li, N. Wang, Q. Liu, GDP spatialization and economic differences in South China based on NPP-VIIRS nighttime light imagery, *Remote Sensing* 9 (7) (2017) 673.
- [222] K. Baugh, F. C. Hsu, C. D. Elvidge, M. Zhizhin, Nighttime lights compositing using the VIIRS day-night band: Preliminary results, *Proceedings of the Asia-Pacific Advanced Network* 35 (2013) 70–86.
- [223] Z. Dai, Y. Hu, G. Zhao, The suitability of different nighttime light data for GDP estimation at different spatial scales and regional levels, *Sustainability* 9 (2) (2017) 305.
- [224] W. Wang, H. Cheng, L. Zhang, Poverty assessment using DMSP/OLS night-time light satellite imagery at a provincial scale in China, *Advances in Space Research* 49 (8) (2012) 1253–1264.
- [225] R. Engstrom, J. Hersh, D. Newhouse, Poverty from space: Using high-resolution satellite imagery for estimating economic well-being, *Policy Research Working Paper No. 8284*, World Bank Group, Washington, D.C., USA (2017).
- [226] R. Sliuzas, G. Mboup, A. de Sherbinin, Report of the expert group meeting on slum identification and mapping, Tech. rep., CIESIN, UN-Habitat, ITC, Enschede, Netherlands (2008).
- [227] H. Rhinane, A. Hilali, A. Berrada, M. Hakdaoui, Detecting slums from SPOT data in Casablanca Morocco using an object based approach, *Journal of Geographic Information System* 3 (3) (2011) 217–224.
- [228] S. Shekhar, Detecting slums from Quick Bird data in Pune using an object oriented approach, *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 39 (8) (2012) 519–524.
- [229] P. Cheng, T. Toutin, Y. Zhang, QuickBird–Geometric correction, data fusion, and automatic DEM extraction, *Earth Observation Magazine* 11 (4) (2003) 14–18.
- [230] D. Kohli, R. Sliuzas, N. Kerle, A. Stein, An ontology of slums for image-based classification, *Computers, Environment and Urban Systems* 36 (2) (2012) 154–163.
- [231] O. Kit, M. Lüdeke, D. Reckien, Texture-based identification of urban slums in Hyderabad, India using remote sensing data, *Applied Geography* 32 (2) (2012) 660–667.
- [232] K. Martinez, J. Cupitt, VIPS—A highly tuned image processing software architecture, in: *IEEE International Conference on Image Processing*, Vol. 2, IEEE Press, 2005, pp. 574–577.
- [233] Y. Malhi, R. M. Román-Cuesta, Analysis of lacunarity and scales of spatial homogeneity in IKONOS images of Amazonian tropical forest canopies, *Remote Sensing of Environment* 112 (5) (2008) 2074–2087.
- [234] O. Kit, M. Lüdeke, Automated detection of slum area change in Hyderabad, India using multitemporal satellite imagery, *ISPRS Journal of Photogrammetry and Remote Sensing* 83 (2013) 130–137.
- [235] J. Canny, A computational approach to edge detection, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 8 (6) (1986) 679–698.
- [236] R. G. vn Gioi, J. Jakubowicz, J. M. Morel, G. Randall, LSD: A line segment detector, *Image Processing on Line* 2 (4) (2012) 35–55.
- [237] O. Gruebner, J. Sachs, A. Nockert, M. Frings, M. M. H. Khan, T. Lakes, P. Hostert, Mapping the slums of Dhaka from 2006 to 2010, *Dataset Papers in Science* 2014 (2014) 172182.
- [238] R. Engstrom, A. Sandborn, Q. Yu, J. Burgdorfer, D. Stow, J. Weeks, J. Graesser, Mapping slums using spatial features in Accra, Ghana, in: *2015 Joint Urban Remote Sensing Event*, IEEE Press, 2015, pp. 1–4.
- [239] D. Kohli, A. Stein, R. Sliuzas, Uncertainty analysis for image interpretations of urban slums, *Computers Environment and Urban Systems*

60 (2016) 37–49.

- [240] M. Kuffer, K. Pfeffer, R. Sliuzas, I. Baud, Extraction of slum areas from VHR imagery using GLCM variance, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 9 (5) (2016) 1830–1840.
- [241] M. Wurm, H. Taubenböck, M. Weigand, A. Schmitt, Slum mapping in polarimetric SAR data using spatial features, *Remote Sensing of Environment* 194 (2017) 190–204.
- [242] A. Schmitt, A. Wendleder, S. Hinz, The Kennaugh element framework for multi-scale, multi-polarized, multi-temporal and multi-frequency SAR image preparation, *ISPRS Journal of Photogrammetry and Remote Sensing* 102 (2015) 122–139.
- [243] M. Kuffer, K. Pfeffer, R. Sliuzas, Slums from space—15 years of slum mapping using remote sensing, *Remote Sensing* 8 (6) (2016) 455.
- [244] R. Mahabir, A. Croitoru, A. T. Crooks, P. Agouris, A. Stefanidis, A critical review of high and very high-resolution remote sensing approaches for detecting and mapping slums: Trends, challenges and emerging opportunities, *Urban Science* 2 (1) (2018) 8.
- [245] M. Woolcock, Social capital and economic development: Toward a theoretical synthesis and policy framework, *Theory and Society* 27 (2) (1998) 151–208.
- [246] P. S. Adler, S.-W. Kwon, Social capital: Prospects for a new concept, *Academy of Management Review* 27 (1) (2002) 17–40.
- [247] M. Granovetter, The impact of social structure on economic outcomes, *Journal of Economic Perspectives* 19 (1) (2005) 33–50.
- [248] N. Eagle, M. Macy, R. Claxton, Network diversity and economic development, *Science* 328 (5981) (2010) 1029–1031.
- [249] R. S. Burt, *Structural Holes: The Social Structure of Competition*, Harvard University Press, New York, NY, USA, 2009.
- [250] A. Rubio, V. Frias-Martinez, E. Frias-Martinez, N. Oliver, Human mobility in advanced and developing economies: A comparative analysis, in: *Proceedings of the AAAI Spring Symposium: Artificial Intelligence for Development*, AI-D'10, AAAI Press, Palo Alto, CA, USA, 2010, pp. 79–84.
- [251] V. Frias-Martinez, J. Virseda, On the relationship between socio-economic factors and cell phone usage, in: *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development*, ICTD'12, ACM Press, New York, NY, USA, 2012, pp. 76–84.
- [252] C. Smith-Clarke, A. Mashhadni, L. Capra, Poverty on the cheap: Estimating poverty maps using aggregated mobile communication networks, in: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI'14, ACM Press, New York, NY, USA, 2014, pp. 511–520.
- [253] S. Šćepanović, I. Mishkovski, P. Hui, J. K. Nurminen, A. Ylä-Jääski, Mobile phone call data as a regional socio-economic proxy indicator, *PLoS ONE* 10 (4) (2015) e0124160.
- [254] Y. Leo, E. Fleury, J. I. Alvarez-Hamelin, C. Sarraute, M. Karsai, Socioeconomic correlations and stratification in social-communication networks, *Journal of the Royal Society Interface* 13 (125) (2016) 20160598.
- [255] Y. Leo, M. Karsai, C. Sarraute, E. Fleury, Correlations of consumption patterns in social-economic networks, in: *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM'16, IEEE Press, 2016, pp. 493–500.
- [256] Z. Cheng, J. Caverlee, K. Lee, You are where you tweet: A content-based approach to geo-locating Twitter users, in: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM'10, ACM Press, New York, NY, USA, 2010, pp. 759–768.
- [257] D. Quercia, J. Ellis, L. Capra, J. Crowcroft, Tracking gross community happiness from tweets, in: *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW'12, ACM Press, New York, NY, USA, 2012, pp. 965–968.
- [258] A. D. I. Kramer, An unobtrusive behavioral model of “gross national happiness”, in: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI'10, ACM Press, New York, NY, USA, 2010, pp. 287–290.
- [259] J. Mahmud, J. Nichols, C. Drews, Where is this tweet from? Inferring home locations of Twitter users, in: *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, ICWSM'12, AAAI Press, Palo Alto, CA, USA, 2012, pp. 511–514.
- [260] D. Jimenez, Dynamically weighted ensemble neural networks for classification, in: *IEEE International Joint Conference on Neural Networks Proceedings*, IEEE Press, 1998, pp. 753–756.
- [261] S. Hasan, X. Zhan, S. V. Ukkusuri, Understanding urban human activity and mobility patterns using large-scale location-based data from online social media, in: *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, UrbComp'13, ACM Press, New York, NY, USA, 2013, p. 6.
- [262] S. Hasan, S. V. Ukkusuri, Urban activity pattern classification using topic models from online geo-location data, *Transportation Research Part C* 44 (2014) 363–381.
- [263] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, *Journal of Machine Learning Research* 3 (2003) 993–1022.
- [264] G. Lansley, P. A. Longley, The geography of Twitter topics in London, *Computers Environment and Urban Systems* 58 (2016) 85–96.
- [265] Q. Huang, D. W. Wong, Activity patterns, socioeconomic status and urban spatial structure: What can social media data tell us?, *International Journal of Geographical Information Science* 30 (9) (2016) 1873–1898.
- [266] J.-H. Liu, J. Wang, J. Shao, T. Zhou, Online social activity reflects economic status, *Physica A: Statistical Mechanics and its Applications* 457 (2016) 581–589.
- [267] S. M. Stigler, Francis Galton's account of the invention of correlation, *Statistical Science* 4 (2) (1989) 73–79.
- [268] J. L. Myers, A. D. Well, *Research Design and Statistical Analysis*, 3rd Edition, Routledge, New York, NY, USA, 2010.
- [269] J. Wang, J. Gao, J.-H. Liu, D. Yang, T. Zhou, Regional economic status inference from information flow and talent mobility, *EPL (Euro-physics Letters)* 125 (6) (2019) 68002.
- [270] X. Yang, J. Gao, J. H. Liu, T. Zhou, Height conditions salary expectations: Evidence from large-scale data in china, *Physica A: Statistical Mechanics and its Applications* 501 (2018) 86–97.
- [271] T. Nguyen, B. K. Szymanski, Using location-based social networks to validate human mobility and relationships models, in: *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, IEEE Press, 2012, pp. 1215–1221.
- [272] B. O. Holzbauer, B. K. Szymanski, T. Nguyen, A. Pentland, Social ties as predictors of economic development, in: *Proceedings of the 12th International Conference and School on Advances in Network Science*, Vol. 9564 of NetSci-X 2016, Springer, New York, NY, USA, 2016, pp. 178–185.
- [273] L. Norbutas, R. Corten, Network structure and economic prosperity in municipalities: A large-scale test of social capital theory using social

- media data, *Social Networks* 52 (2018) 120–134.
- [274] S. Scellato, C. Mascolo, M. Musolesi, V. Latora, Distance matters: Geo-social metrics for online social networks, in: Proceedings of the 3rd Wonference on Online Social Networks, WOSN’10, USENIX Association, Berkeley, CA, USA, 2010, p. 8.
- [275] R. Guimera, M. Sales-Pardo, L. A. N. Amaral, Modularity from fluctuations in random graphs and complex networks, *Physical Review E* 70 (2) (2004) 025101.
- [276] M. E. Newman, Modularity and community structure in networks, *Proceedings of the National Academy of Sciences of the United States of America* 103 (23) (2006) 8577–8582.
- [277] A. Venerandi, G. Quattrone, L. Capra, D. Quercia, D. Saez-Trumper, Measuring urban deprivation from user generated content, in: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW’15, ACM Press, New York, NY, USA, 2015, pp. 254–264.
- [278] B. T. van Zanten, D. B. van Berkel, R. K. Meentemeyer, J. W. Smith, K. F. Tieskens, P. H. Verburg, Continental-scale quantification of landscape values using social media data, *Proceedings of the National Academy of Sciences of the United States of America* 113 (46) (2016) 12974–12979.
- [279] C. Cortes, V. Vapnik, Support-vector networks, *Machine Learning* 20 (3) (1995) 273–297.
- [280] A. J. Smola, B. Schölkopf, A tutorial on support vector regression, *Statistics and Computing* 14 (3) (2004) 199–222.
- [281] J. Gao, B. Jun, A. S. Pentland, T. Zhou, C. A. Hidalgo, Collective learning in China’s regional economic development, arXiv:1703.01369, 2017.
- [282] J. Gao, T. Zhou, Quantifying China’s regional economic complexity, *Physica A: Statistical Mechanics and its Applications* 492 (2018) 1591–1603.
- [283] Q. Guo, C. He, Production space and regional industrial evolution in China, *GeoJournal* 82 (2) (2017) 379–396.
- [284] R. Boschma, S. Iammarino, Related variety, trade linkages, and regional growth in Italy, *Economic Geography* 85 (3) (2009) 289–311.
- [285] R. Boschma, K. Frenken, Technological relatedness and regional branching, in: H. Bathelt, M. P. Feldman, D. F. Kogler (Eds.), *Beyond Territory. Dynamic Geographies of Knowledge Creation, Diffusion and Innovation*, Routledge, London, UK, 2012, pp. 64–81.
- [286] R. Boschma, A. Minondo, M. Navarro, Related variety and regional growth in Spain, *Papers in Regional Science* 91 (2) (2012) 241–256.
- [287] C. Castaldi, K. Frenken, B. Los, Related variety, unrelated variety and technological breakthroughs: An analysis of US state-level patenting, *Regional Studies* 49 (5) (2015) 767–781.
- [288] R. Boschma, P. A. Balland, D. F. Kogler, Relatedness and technological change in cities: The rise and fall of technological knowledge in US metropolitan areas from 1981 to 2010, *Industrial and Corporate Change* 24 (1) (2015) 223–250.
- [289] P. A. Balland, R. Boschma, K. Frenken, Proximity and innovation: From statics to dynamics, *Regional Studies* 49 (6) (2015) 907–920.
- [290] D. Acemoglu, U. Akcigit, W. R. Kerr, Innovation network, *Proceedings of the National Academy of Sciences of the United States of America* 113 (41) (2016) 11483–11488.
- [291] C. A. Hidalgo, P.-A. Balland, R. Boschma, M. Delgado, M. Feldman, K. Frenken, E. Glaeser, C. He, D. F. Kogler, A. Morrison, F. Neffke, D. Rigby, S. Stern, S. Zheng, S. Zhu, The principle of relatedness, in: Proceedings of the Ninth International Conference on Complex Systems, ICCS 2018, Springer, Cham, Switzerland, 2018, pp. 451–457.
- [292] B. Jun, A. Alshamsi, J. Gao, C. A. Hidalgo, Relatedness, knowledge diffusion, and the evolution of bilateral trade, arXiv:1709.05392, 2017.
- [293] R. Boschma, Relatedness as driver of regional diversification: A research agenda, *Regional Studies* 51 (3) (2017) 351–364.
- [294] M. Davids, K. Frenken, Proximity, knowledge base and the innovation process: Towards an integrated framework, *Regional Studies* 52 (1) (2018) 23–34.
- [295] R. Boschma, A. Minondo, M. Navarro, The emergence of new industries at the regional level in Spain: A proximity approach based on product relatedness, *Economic Geography* 89 (1) (2013) 29–51.
- [296] W. Keller, Geographic localization of international technology diffusion, *American Economic Review* 92 (1) (2002) 120–142.
- [297] D. Bahar, R. Hausmann, C. A. Hidalgo, Neighbors and the evolution of the comparative advantage of nations: Evidence of international knowledge diffusion?, *Journal of International Economics* 92 (1) (2014) 111–123.
- [298] C. Lawson, E. Lorenz, Collective learning, tacit knowledge and regional innovative capacity, *Regional Studies* 33 (4) (1999) 305–317.
- [299] Z. J. Acs, C. Armington, T. Zhang, The determinants of new-firm survival across regional economies: The role of human capital stock and knowledge spillover, *Papers in Regional Science* 86 (3) (2007) 367–391.
- [300] T. J. Holmes, The diffusion of Wal-Mart and economies of density, *Econometrica* 79 (1) (2011) 253–302.
- [301] T. Broekel, R. Boschma, Knowledge networks in the Dutch aviation industry: The proximity paradox, *Journal of Economic Geography* 12 (2) (2012) 409–433.
- [302] C. Jara-Figueroa, B. Jun, E. L. Glaeser, C. A. Hidalgo, The role of industry-specific, occupation-specific, and location-specific knowledge in the growth and survival of new firms, *Proceedings of the National Academy of Sciences of the United States of America* 115 (50) (2018) 12646–12653.
- [303] A. Alabdulkareem, M. Frank, L. Sun, B. AlShebli, C. A. Hidalgo, I. Rahwan, Unpacking the polarization of workplace skills, *Science Advances* 4 (7) (2018) eaao6030.
- [304] T. Broekel, R. Boschma, The cognitive and geographical structure of knowledge links and how they influence firms’ innovation performance, *Regional Statistics* 6 (2) (2016) 3–26.
- [305] R. Boschma, V. Martín, A. Minondo, Neighbour regions as the source of new industries, *Papers in Regional Science* 96 (2) (2017) 227–245.
- [306] J. Gao, B. Jun, T. Zhou, C. A. Hidalgo, Revealing and maximizing the collective learning effects in Brazilian industrial diversification, unpublished (2017).
- [307] K. J. Arrow, Classificatory notes on the production and transmission of technological knowledge, *American Economic Review* 59 (2) (1969) 29–35.
- [308] S. Zheng, M. E. Kahn, China’s bullet trains facilitate market integration and mitigate the cost of megacity growth, *Proceedings of the National Academy of Sciences of the United States of America* 110 (14) (2013) E1248–E1253.
- [309] X. Li, B. Huang, R. Li, Y. Zhang, Exploring the impact of high speed railways on the spatial redistribution of economic activities—Yangtze River Delta urban agglomeration as a case study, *Journal of Transport Geography* 57 (2016) 194–206.

- [310] X. Ke, H. Chen, Y. Hong, H. Cheng, Do China's high-speed-rail projects promote local economy?—New evidence from a panel data approach, *China Economic Review* 44 (2017) 203–226.
- [311] Y. Qin, "No county left behind?" The distributional impact of high-speed rail upgrades in China, *Journal of Economic Geography* 17 (3) (2017) 489–520.
- [312] Y. S. Cheng, B. P. Y. Loo, R. Vickerman, High-speed rail networks, economic integration and regional specialisation in China and Europe, *Travel Behaviour and Society* 2 (1) (2015) 1–14.
- [313] R. Vickerman, Can high-speed rail have a transformative effect on the economy?, *Transport Policy* 62 (2018) 31–37.
- [314] G. M. Ahlfeldt, A. Feddersen, From periphery to core: Measuring agglomeration effects using high-speed rail, *Journal of Economic Geography* 18 (2) (2018) 355–390.
- [315] C. Catalini, C. Fons-Rosen, P. Gaulé, Did cheaper flights change the direction of science?, IZA Discussion Papers No. 9897, IZA Institute of Labor Economics, Bonn, Germany (2016).
- [316] F. Neffke, M. S. Henning, Revealed relatedness: Mapping industry space, *Papers in Evolutionary Economic Geography* No. 08.19, Utrecht University, Utrecht, Netherlands (2008).
- [317] F. M. Neffke, M. Henning, R. Boschma, The impact of aging and technological relatedness on agglomeration externalities: A survival analysis, *Journal of Economic Geography* 12 (2) (2012) 485–517.
- [318] F. Neffke, M. Henning, Skill relatedness and firm diversification, *Strategic Management Journal* 34 (3) (2013) 297–316.
- [319] C. He, F. Pan, T. Chen, Research progress of industrial geography in China, *Journal of Geographical Sciences* 26 (8) (2016) 1057–1066.
- [320] A. Howell, C. He, R. Yang, C. C. Fand, Agglomeration,(un)-related variety and new firm survival in China: Do local subsidies matter?, *Papers in Regional Science* 97 (3) (2018) 485–500.
- [321] C. He, Q. Guo, D. Rigby, What sustains larger firms? Evidence from Chinese manufacturing industries, *Annals of Regional Science* 58 (2) (2017) 275–300.
- [322] C. He, Y. Yan, D. Rigby, Regional industrial evolution in China, *Papers in Regional Science* 97 (2) (2018) 173–198.
- [323] J. Gao, Maximizing the collective learning effects in regional economic development, in: 2017 14th International Computer Conference on Wavelet Active Media Technology and Information Processing, IEEE Press, 2017, pp. 337–341.
- [324] D. J. Watts, A simple model of global cascades on random networks, *Proceedings of the National Academy of Sciences of the United States of America* 99 (9) (2002) 5766–5771.
- [325] J. Gao, T. Zhou, Y. Hu, Bootstrap percolation on spatial networks, *Scientific Reports* 5 (2015) 14662.
- [326] A. Alshamsi, F. L. Pinheiro, C. A. Hidalgo, Optimal diversification strategies in the networks of related products and of related research areas, *Nature Communications* 9 (2018) 1328.
- [327] S. Zhu, C. He, Y. Zhou, How to jump further and catch up? Path-breaking in an uneven industry space, *Journal of Economic Geography* 17 (3) (2017) 521–545.
- [328] R. Boschma, L. Coenen, K. Frenken, B. Truffer, Towards a theory of regional diversification: Combining insights from Evolutionary Economic Geography and Transition Studies, *Regional Studies* 51 (1) (2017) 31–45.
- [329] F. L. Pinheiro, A. Alshamsi, D. Hartmann, R. Boschma, C. A. Hidalgo, Shooting low or high: Do countries benefit from entering unrelated activities?, arXiv:1801.05352, 2018.
- [330] X. Gabaix, Zipf's law for cities: An explanation, *Quarterly Journal of Economics* 114 (3) (1999) 739–767.
- [331] A. Clauset, C. R. Shalizi, M. E. J. Newman, Power-law distributions in empirical data, *SIAM Review* 51 (4) (2009) 661–703.
- [332] C. Small, C. D. Elvidge, D. Balk, M. Montgomery, Spatial scaling of stable night lights, *Remote Sensing of Environment* 115 (2) (2011) 269–280.
- [333] L. M. Bettencourt, J. Lobo, D. Helbing, C. Kühnert, G. B. West, Growth, innovation, scaling, and the pace of life in cities, *Proceedings of the National Academy of Sciences of the United States of America* 104 (17) (2007) 7301–7306.
- [334] E. Arcarate, E. Hatna, P. Ferguson, H. Youn, A. Johansson, M. Batty, Constructing cities, deconstructing scaling laws, *Journal of the Royal Society Interface* 12 (102) (2015) 20140745.
- [335] R. Louf, M. Barthélémy, How congestion shapes cities: From mobility patterns to scaling, *Scientific Reports* 4 (2014) 5561.
- [336] E. A. Oliveira, J. S. Andrade Jr., H. A. Makse, Large cities are less green, *Scientific Reports* 4 (2014) 4235.
- [337] J. P. Delong, O. Burger, Socio-economic instability and the scaling of energy use with population size, *PLoS ONE* 10 (6) (2015) e0130547.
- [338] H. Samaniego, M. E. Moses, Cities as organisms: Allometric scaling of urban road networks, *Journal of Transport and Land Use* 1 (1) (2008) 21–39.
- [339] M. Batty, A theory of city size, *Science* 340 (6139) (2013) 1418–1419.
- [340] R. Louf, C. Roth, M. Barthélémy, Scaling in transportation networks, *PLoS ONE* 9 (7) (2014) e102007.
- [341] A. P. Masucci, E. Arcarate, E. Hatna, K. Stanilov, M. Batty, On the problem of boundaries and scaling for urban street networks, *Journal of the Royal Society Interface* 12 (111) (2015) 20150763.
- [342] L. G. A. Alves, H. V. Ribeiro, R. S. Mendes, Scaling laws in the dynamics of crime growth rate, *Physica A: Statistical Mechanics and its Applications* 392 (11) (2013) 2672–2679.
- [343] S. Banerjee, P. van Hentenryck, M. Cebrian, Competitive dynamics between criminals and law enforcement explains the super-linear scaling of crime in cities, *Palgrave Communications* 1 (2015) 15022.
- [344] Q. S. Hanley, L. Dan, H. V. Ribeiro, Rural to urban population density scaling of crime and property transactions in English and Welsh Parliamentary Constituencies, *PLoS ONE* 11 (2) (2016) e0149546.
- [345] M. Oliveira, C. Bastos-Filho, R. Menezes, The scaling of crime concentration in cities, *PLoS ONE* 12 (8) (2017) e0183110.
- [346] A. F. J. van Raan, G. van der Meulen, W. Goedhart, Urban scaling of cities in the Netherlands, *PLoS ONE* 11 (1) (2016) e0146775.
- [347] L. M. A. Bettencourt, L. Jose, Urban scaling in Europe, *Journal of the Royal Society Interface* 13 (116) (2016) 20160005.
- [348] W. Pan, G. Ghoshal, C. Krumme, M. Cebrian, A. Pentland, Urban characteristics attributable to density-driven tie formation, *Nature Communications* 4 (2013) 1961.
- [349] T. Louail, M. Lenormand, O. G. C. Ros, M. Picornell, R. Herranz, E. Frias-Martinez, J. J. Ramasco, M. Barthélémy, From mobile phone data to the spatial structure of cities, *Scientific Reports* 4 (2014) 5276.

- [350] M. Schlapfer, L. M. A. Bettencourt, S. Grauwin, M. Raschke, R. Claxton, Z. Smoreda, G. B. West, C. Ratti, The scaling of human interactions with city size, *Journal of the Royal Society Interface* 11 (98) (2014) 20130789.
- [351] J. C. Leitão, J. M. Miotto, M. Gerlach, E. G. Altmann, Is this scaling nonlinear?, *Royal Society Open Science* 3 (7) (2016) 150649.
- [352] S. Arbesman, J. M. Kleinberg, S. H. Strogatz, Superlinear scaling for innovation in cities, *Physical Review E* 79 (2) (2009) 016115.
- [353] A. Gomez-Lievano, H. Youn, L. M. Bettencourt, The statistics of urban scaling and their connection to Zipf's law, *PLoS ONE* 7 (7) (2012) e40393.
- [354] L. M. Bettencourt, The origins of scaling in cities, *Science* 340 (6139) (2013) 1438–1441.
- [355] K. Yakubo, Y. Saijo, D. Korošak, Superlinear and sublinear urban scaling in geographical networks modeling cities, *Physical Review E* 90 (2) (2014) 022803.
- [356] A. Gomez-Lievano, O. Patterson-Lomba, R. Hausmann, Explaining the prevalence, scaling and variance of urban phenomena, *Nature Human Behaviour* 1 (2016) 12.
- [357] F. L. Ribeiro, J. Meirelles, F. F. Ferreira, C. R. Neto, A model of urban scaling laws based on distance dependent interactions, *Royal Society Open Science* 4 (3) (2017) 160926.
- [358] R. Li, L. Dong, J. Zhang, X. Wang, W.-X. Wang, Z. Di, H. E. Stanley, Simple spatial scaling rules behind complex cities, *Nature Communications* 8 (2017) 1841.
- [359] L. M. Bettencourt, J. Lobo, D. Strumsky, G. B. West, Urban scaling and its deviations: Revealing the structure of wealth, innovation and crime across cities, *PLoS ONE* 5 (11) (2010) e13541.
- [360] M. Batty, L. March, The method of residues in urban modelling, *Environment and Planning A* 8 (2) (2008) 189–214.
- [361] J. Lobo, L. M. A. Bettencourt, D. Strumsky, G. B. West, Urban scaling and the production function for cities, *PLoS ONE* 8 (3) (2013) e58407.
- [362] H. Youn, L. M. Bettencourt, J. Lobo, D. Strumsky, H. Samaniego, G. B. West, Scaling and universality in urban economic diversification, *Journal of the Royal Society Interface* 13 (114) (2016) 20150957.
- [363] P. C. Sutton, A scale-adjusted measure of “urban sprawl” using nighttime satellite imagery, *Remote Sensing of Environment* 86 (3) (2003) 353–369.
- [364] B. Pandey, P. Joshi, K. C. Seto, Monitoring urbanization dynamics in India using DMSP/OLS night time lights and SPOT-VGT data, *International Journal of Applied Earth Observation and Geoinformation* 23 (2013) 49–61.
- [365] O. Hagolle, A. Lobo, P. Maisongrande, F. Cabot, B. Duchemin, A. D. Pereyra, Quality assessment and improvement of temporally composited products of remotely sensed imagery by combination of VEGETATION 1 and 2 images, *Remote Sensing of Environment* 94 (2) (2005) 172–186.
- [366] X. Cao, J. Chen, H. Imura, O. Higashi, A SVM-based method to extract urban areas from DMSP-OLS and SPOT VGT data, *Remote Sensing of Environment* 113 (10) (2009) 2205–2209.
- [367] Q. Zhang, K. C. Seto, Mapping urbanization dynamics at regional and global scales using multi-temporal DMSP/OLS nighttime light data, *Remote Sensing of Environment* 115 (9) (2011) 2320–2329.
- [368] S. Frolking, T. Milliman, K. C. Seto, M. A. Friedl, A global fingerprint of macro-scale changes in urban structure from 1999 to 2009, *Environmental Research Letters* 8 (2) (2013) 024004.
- [369] D. G. Long, M. R. Drinkwater, B. Holt, S. Saatchi, C. Bertoia, Global ice and land climate studies using scatterometer image data, *EOS Transactions American Geophysical Union* 82 (43) (2013) 503–503.
- [370] Q. Li, L. Lu, Q. Weng, Y. Xie, H. Guo, Monitoring urban dynamics in the Southeast U.S.A. using time-series DMSP/OLS nightlight imagery, *Remote Sensing* 8 (7) (2016) 578.
- [371] X. Huang, A. Schneider, M. A. Friedl, Mapping sub-pixel urban expansion in China using MODIS and DMSP/OLS nighttime lights, *Remote Sensing of Environment* 175 (2016) 92–108.
- [372] A. Schneider, M. A. Friedl, D. Potere, Mapping global urban areas using MODIS 500-m data: New methods and datasets based on ‘urban ecoregions’, *Remote Sensing of Environment* 114 (8) (2010) 1733–1746.
- [373] Z. Liu, C. He, J. Wu, General spatiotemporal patterns of urbanization: An examination of 16 world cities, *Sustainability* 8 (1) (2016) 41.
- [374] G. Chi, J.-C. Thill, D. Tong, L. Shi, Y. Liu, Uncovering regional characteristics from mobile phone data: A network science approach, *Papers in Regional Science* 95 (3) (2016) 613–631.
- [375] L. C. Freeman, A set of measures of centrality based on betweenness, *Sociometry* 40 (1) (1977) 35–41.
- [376] M. Girvan, M. E. J. Newman, Community structure in social and biological networks, *Proceedings of the National Academy of Sciences of the United States of America* 99 (12) (2002) 7821–7826.
- [377] J. L. Toole, M. Ulm, M. C. González, D. Bauer, Inferring land use from mobile phone activity, in: *Proceedings of the ACM SIGKDD International Workshop on Urban Computing, UrbComp'12*, ACM Press, New York, NY, USA, 2012, pp. 1–8.
- [378] N. J. Yuan, Y. Zheng, X. Xie, Discovering regions of different functions in a city using human mobility and POIs, in: *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'12*, ACM Press, New York, NY, USA, 2012, pp. 186–194.
- [379] N. J. Yuan, Y. Zheng, X. Xie, Y. Wang, K. Zheng, H. Xiong, Discovering urban functional zones using latent activity trajectories, *IEEE Transactions on Knowledge and Data Engineering* 27 (3) (2015) 712–725.
- [380] V. Frias-Martinez, V. Soto, H. Hohwald, E. Frias-Martinez, Characterizing urban landscapes using geolocated tweets, in: *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust, SOCIALCOM-PASSAT'12*, IEEE Press, 2012, pp. 239–248.
- [381] V. Frias-Martinez, E. Frias-Martinez, Spectral clustering for sensing urban land use using Twitter activity, *Engineering Applications of Artificial Intelligence* 35 (10) (2014) 237–245.
- [382] A. Lloyd, J. Cheshire, Deriving retail centre locations and catchments from geo-tagged Twitter data, *Computers Environment and Urban Systems* 61 (2017) 108–118.
- [383] A. Soliman, K. Soltani, J. Yin, A. Padmanabhan, S. Wang, Social sensing of urban land use based on analysis of Twitter users’ mobility patterns, *PLoS ONE* 12 (7) (2017) 0181657.

- [384] Y. Shen, K. Karimi, Urban function connectivity: Characterisation of functional urban streets with social media check-in data, *Cities* 55 (2016) 9–21.
- [385] Y. Zhi, H. Li, D. Wang, M. Deng, S. Wang, J. Gao, Z. Duan, Y. Liu, Latent spatio-temporal activity structures: A new approach to inferring intra-urban functional regions via social media check-in data, *Geo-spatial Information Science* 19 (2) (2016) 94–105.
- [386] S. Jiang, A. Alves, F. Rodrigues, J. Ferreira Jr., F. C. Pereira, Mining point-of-interest data from social networks for urban land use classification and disaggregation, *Computers Environment and Urban Systems* 53 (2015) 36–46.
- [387] Y. Hu, S. Gao, K. Janowicz, B. Yu, W. Li, S. Prasad, Extracting and understanding urban areas of interest using geotagged photos, *Computers Environment and Urban Systems* 54 (2015) 240–254.
- [388] M. Ester, H. P. Kriegel, X. Xu, A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise, in: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD’96, AAAI Press, Palo Alto, CA, USA, 1996, pp. 226–231.
- [389] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, *Information Processing and Management* 24 (5) (1988) 513–523.
- [390] G. S. Sirmans, D. A. Macpherson, E. N. Zietz, The composition of hedonic pricing models, *Journal of Real Estate Literature* 13 (1) (2005) 3–43.
- [391] L. Xin, S. Zheng, Spatial analysis and spatial house price index construction: Evidence from Chengdu housing market, in: Proceedings of the 17th International Symposium on Advancement of Construction Management and Real Estate, Springer, Berlin, Heidelberg, 2014, pp. 1207–1217.
- [392] Y. Zhang, S. Zheng, Y. Song, Y. Zhong, The spillover effect of urban village removal on nearby home values in Beijing, *Growth and Change* 47 (1) (2016) 9–31.
- [393] S. Zheng, Y. Xu, X. Zhang, R. Wang, Transit development, consumer amenities and home values: Evidence from Beijing’s subway neighborhoods, *Journal of Housing Economics* 33 (2016) 22–33.
- [394] S. Zheng, W. Sun, M. E. Kahn, Investor confidence as a determinant of China’s urban housing market dynamics, *Real Estate Economics* 44 (4) (2016) 814–845.
- [395] A. J. Bency, S. Rallapalli, R. K. Ganti, M. Srivatsa, B. Manjunath, Beyond spatial auto-regressive models: Predicting housing prices with satellite imagery, in: 2017 IEEE Winter Conference on Applications of Computer Vision, IEEE Press, 2017, pp. 320–329.
- [396] D. Anguelov, C. Dulong, D. Filip, C. Frueh, S. Lafon, R. Lyon, A. Ogale, L. Vincent, J. Weaver, Google Street View: Capturing the world at street level, *Computer* 43 (6) (2010) 32–38.
- [397] A. G. Rundle, M. D. M. Bader, C. A. Richards, K. M. Neckerman, J. O. Teitler, Using Google Street View to audit neighborhood environments, *American Journal of Preventive Medicine* 40 (1) (2011) 94–100.
- [398] P. Salesses, K. Schechtner, C. A. Hidalgo, The collaborative image of the city: Mapping the inequality of urban perception, *PLoS ONE* 8 (7) (2013) e68400.
- [399] V. Ordonez, T. L. Berg, Learning high-level judgments of urban perception, in: Proceedings of the 13th European Conference on Computer Vision, ECCV 2014, Springer, Cham, Switzerland, 2014, pp. 494–510.
- [400] N. Naik, J. Philipoom, R. Raskar, C. A. Hidalgo, Streetscore—predicting the perceived safety of one million streetscapes, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPRW’14, IEEE Press, 2014, pp. 793–799.
- [401] R. Herbrich, T. Minka, T. Graepel, TrueSkill™: A Bayesian skill rating system, in: Proceedings of the 19th International Conference on Neural Information Processing Systems, NIPS’06, MIT Press, Cambridge, MA, USA, 2006, pp. 569–576.
- [402] B. Schölkopf, A. J. Smola, R. Williamson, P. Bartlett, New support vector algorithms, *Neural Computation* 12 (5) (2000) 1207–1245.
- [403] L. Porzi, S. Rota Bulò, B. Lepri, E. Ricci, Predicting and understanding urban perception with convolutional neural networks, in: Proceedings of the 23rd ACM International Conference on Multimedia, MM’15, ACM Press, New York, NY, USA, 2015, pp. 139–148.
- [404] D. Quercia, N. K. O’Hare, H. Cramer, Aesthetic capital: What makes london look beautiful, quiet, and happy?, in: Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW’14, ACM Press, New York, NY, USA, 2014, pp. 945–955.
- [405] S. M. Arietta, A. A. Efros, R. Ramamoothri, M. Agrawala, City forensics: Using visual elements to predict non-visual city attributes, *IEEE Transactions on Visualization and Computer Graphics* 20 (12) (2014) 2624–2633.
- [406] B. E. Boser, I. M. Guyon, V. N. Vapnik, A training algorithm for optimal margin classifiers, in: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, COLT’92, ACM Press, New York, NY, USA, 1992, pp. 144–152.
- [407] C. I. Seresinhe, T. Preis, H. S. Moat, Using deep learning to quantify the beauty of outdoor places, *Royal Society Open Science* 4 (7) (2017) 170170.
- [408] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, A. Oliva, Learning deep features for scene recognition using places database, in: Proceedings of the 27th International Conference on Neural Information Processing Systems, NIPS’14, MIT Press, Cambridge, MA, USA, 2014, pp. 487–495.
- [409] A. Dubey, N. Naik, D. Parikh, R. Raskar, C. A. Hidalgo, Deep learning the city: Quantifying urban perception at a global scale, in: Proceedings of the 14th European Conference on Computer Vision, ECCV 2016, Springer, Cham, Switzerland, 2016, pp. 196–212.
- [410] A. Albert, J. Kaur, M. C. González, Using convolutional networks and satellite imagery to identify patterns in urban environments at a large scale, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD’17, ACM Press, New York, NY, USA, 2017, pp. 1357–1366.
- [411] L. Tracewski, L. Bastin, C. C. Fonte, Repurposing a deep learning network to filter and classify volunteered photographs for land cover and land use characterization, *Geo-spatial Information Science* 20 (3) (2017) 252–268.
- [412] S. Lefévre, D. Tuia, J. D. Wegner, T. Produt, A. S. Nassar, Towards seamless multi-view scene analysis from satellite to street-level, *Proceedings of the IEEE* 105 (10) (2017) 1884–1899.
- [413] M. Guillen, F. Verdú, O. Portolés, A. Castelló, Happiness is greater in natural environments, *Global Environmental Change* 23 (5) (2013) 992–1000.
- [414] M. de Nadai, R. L. Vieriu, G. Zen, S. Dragicevic, N. Naik, M. Caraviello, C. A. Hidalgo, N. Sebe, B. Lepri, Are safer looking neighborhoods

- more lively?: A multimodal investigation into urban life, in: Proceedings of the 2016 ACM on Multimedia Conference, MM'16, ACM Press, New York, NY, USA, 2016, pp. 1127–1135.
- [415] C. Harvey, L. Aultmanhall, Measuring urban streetscapes for livability: A review of approaches, *Professional Geographer* 68 (1) (2016) 149–158.
- [416] N. Naik, S. D. Kominers, R. Raskar, E. L. Glaeser, C. A. Hidalgo, Computer vision uncovers predictors of physical urban change, *Proceedings of the National Academy of Sciences of the United States of America* 114 (29) (2017) 7571–7576.
- [417] E. W. Burgess, The growth of the city, in: R. E. Park, E. W. Burgess, R. D. McKenzie (Eds.), *The City*, University of Chicago Press, Chicago, IL, USA, 1925, pp. 47–62.
- [418] T. C. Schelling, Models of segregation, *American Economic Review* 59 (2) (1969) 488–493.
- [419] E. L. Glaeser, H. Kim, M. Luca, Nowcasting gentrification: Using Yelp data to quantify neighborhood change, *AEA Papers and Proceedings* 108 (2018) 77–82.
- [420] C. Brelsford, T. Martin, J. Hand, L. M. Bettencourt, Toward cities without slums: Topology and the spatial evolution of neighborhoods, *Science Advances* 4 (8) (2018) eaar4644.
- [421] A. Venerandi, G. Quattrone, L. Capra, A scalable method to quantify the relationship between urban form and socio-economic indexes, *EPJ Data Science* 7 (2018) 4.
- [422] Y. Zheng, L. Capra, O. Wolfson, H. Yang, Urban computing: Concepts, methodologies, and applications, *ACM Transactions on Intelligent Systems and Technology* 5 (3) (2014) 38.
- [423] Y. Zheng, *Urban Computing*, MIT Press, Cambridge, MA, USA, 2019.
- [424] F. Calabrese, L. Ferrari, V. D. Blondel, Urban sensing using mobile phone network data: A survey of research, *ACM Computing Surveys* 47 (2) (2015) 25.
- [425] E. L. Glaeser, S. D. Kominers, M. Luca, N. Naik, Big data and big cities: The promises and limitations of improved measures of urban life, *Economic Inquiry* 56 (1) (2018) 114–137.
- [426] J. Jacobs, *The Death and Life of Great American Cities*, Random House, New York, NY, USA, 1961.
- [427] H. Sung, S. Lee, S. Cheon, Operationalizing Jane Jacobs's urban design theory: Empirical verification from the great city of Seoul, Korea, *Journal of Planning Education and Research* 35 (2) (2015) 117–130.
- [428] M. de Nadai, J. Staiano, R. Larcher, N. Sebe, D. Quercia, B. Lepri, The death and life of great italian cities: A mobile phone data perspective, in: Proceedings of the 25th International Conference on World Wide Web, WWW'16, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 2016, pp. 413–423.
- [429] T. Shelton, A. Poorthuis, M. Zook, Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information, *Landscape and Urban Planning* 142 (2015) 198–211.
- [430] D. Wang, F. Li, Y. Chai, Activity spaces and sociospatial segregation in Beijing, *Urban Geography* 33 (2) (2012) 256–277.
- [431] N. M. Yip, R. Forrest, X. Shi, Exploring segregation and mobilities: Application of an activity tracking app on mobile phone, *Cities* 59 (2016) 156–163.
- [432] R. Louf, M. Barthélémy, Patterns of residential segregation, *PLoS ONE* 11 (6) (2016) e0157476.
- [433] J. Hu, Q.-M. Zhang, T. Zhou, Segregation in religion networks, *EPJ Data Science* 8 (2019) 6.
- [434] T. C. Schelling, Dynamic models of segregation, *Journal of Mathematical Sociology* 1 (2) (1971) 143–186.
- [435] A. Sahasranaman, H. J. Jensen, Dynamics of transformation from segregation to mixed wealth cities, *PLoS ONE* 11 (11) (2016) e0166960.
- [436] A. Sahasranaman, H. J. Jensen, Cooperative dynamics of neighborhood economic status in cities, *PLoS ONE* 12 (8) (2017) e0183468.
- [437] T. Gebru, J. Krause, Y. Wang, D. Chen, D. Jia, E. L. Aiden, F.-F. Li, Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States, *Proceedings of the National Academy of Sciences of the United States of America* 114 (50) (2017) 13108–13113.
- [438] P. F. Felzenszwalb, R. B. Girshick, D. Mcallester, D. Ramanan, Object detection with discriminatively trained part-based models, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32 (9) (2010) 1627–1645.
- [439] J. Mycielski, W. Trzeciakowski, Optimization of size and location of service stations, *Journal of Regional Science* 5 (1) (1963) 59–68.
- [440] G. E. Stephan, Territorial subdivision, *Social Forces* 63 (1) (1984) 145–159.
- [441] G. E. Stephan, The distribution of service establishments, *Journal of Regional Science* 28 (1) (1988) 29–40.
- [442] M. T. Gastner, M. E. J. Newman, Optimal design of spatial distribution networks, *Physical Review E* 74 (1) (2006) 016117.
- [443] J. Um, S.-W. Son, S.-I. Lee, H. Jeong, B. J. Kim, Scaling laws between population and facility densities, *Proceedings of the National Academy of Sciences of the United States of America* 106 (34) (2009) 14236–14240.
- [444] D. Karamshuk, A. Noulas, S. Scellato, V. Nicosia, C. Mascolo, Geo-spotting: Mining online location-based services for optimal retail store placement, in: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'13, ACM Press, New York, NY, USA, 2013, pp. 793–801.
- [445] C. A. Hidalgo, E. E. Castañer, The amenity space and the evolution of neighborhoods, arXiv:1509.02868, 2015.
- [446] L. Lü, M. Medo, H. Y. Chi, Y.-C. Zhang, Z.-K. Zhang, T. Zhou, Recommender systems, *Physics Reports* 519 (1) (2012) 1–49.
- [447] T. Louail, M. Lenormand, J. M. Arias, J. J. Ramasco, Crowdsourcing the Robin Hood effect in cities, *Applied Network Science* 2 (2017) 11.
- [448] A. Poddar, J. Foreman, S. Banerjee, P. S. Ellen, Exploring the Robin Hood effect: Moral profiteering motives for purchasing counterfeit products, *Journal of Business Research* 65 (10) (2012) 1500–1506.
- [449] J. Donner, The use of mobile phones by microentrepreneurs in Kigali, Rwanda: Changes to social and business networks, *Information Technologies and International Development* 3 (2) (2006) 2–19.
- [450] V. Soto, V. Frias-Martinez, J. Virseda, E. Frias-Martinez, Prediction of socioeconomic levels using cell phone records, in: Proceedings of the 19th International Conference on User Modeling, Adaption, and Personalization, UMAP'11, Springer, Berlin, Heidelberg, 2011, pp. 377–388.
- [451] J. Blumenstock, N. Eagle, Mobile divides: Gender, socioeconomic status, and mobile phone use in Rwanda, in: Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development, ICTD'10, ACM Press, New York, NY, USA, 2010, p. 6.

- [452] J. E. Blumenstock, D. Gillick, N. Eagle, Who's calling? Demographics of mobile phone use in Rwanda, in: Proceedings of the 2010 AAAI Spring Symposium Series, AAAI Press, Palo Alto, CA, USA, 2010, pp. 116–117.
- [453] A. Wesolowski, N. Eagle, A. M. Noor, R. W. Snow, C. O. Buckee, Heterogeneous mobile phone ownership and usage patterns in Kenya, *PLoS ONE* 7 (4) (2012) e35319.
- [454] J. E. Blumenstock, Calling for better measurement: Estimating an individual's wealth and well-being from mobile phone transaction records, in: Proceedings of the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Workshop on Data Science for Social Good, KDD'14, ACM Press, New York, NY, USA, 2014, pp. 1–6.
- [455] R. R. Agarwal, C. C. Lin, K. T. Chen, V. K. Singh, Predicting financial trouble using call data—On social capital, phone logs, and financial trouble, *PLoS ONE* 13 (2) (2018) e0191863.
- [456] P. Sundsøy, J. Bjelland, B. Reme, A. Iqbal, E. Jahani, Deep learning applied to mobile phone data for individual income classification, in: Proceedings of the 2016 International Conference on Artificial Intelligence: Technologies and Applications, ICAITA 2016, Atlantis Press, Oxford, UK, 2016, pp. 96–99.
- [457] D. Björkegren, D. Grissen, Behavior revealed in mobile phone usage predicts loan repayment, *arXiv:1712.05840*, 2017.
- [458] S. Prina, Banking the poor via savings accounts: Evidence from a field experiment, *Journal of Development Economics* 115 (2015) 16–31.
- [459] X. Dong, E. Jahani, A. Morales-Guzman, B. Bozkaya, B. Lepri, A. S. Pentland, Purchase patterns, socioeconomic status, and political inclination, in: Proceedings of the 2nd Annual International Conference on Computational Social Science, IC2S2 2016, Northwestern University, Evanston, IL, USA, 2016, pp. 1–5.
- [460] B. Hashemian, E. Massaro, I. Bojic, J. M. Arias, S. Sobolevsky, C. Ratti, Socioeconomic characterization of regions through the lens of individual financial transactions, *PLoS ONE* 12 (11) (2017) e0187031.
- [461] S. Sobolevsky, E. Massaro, I. Bojic, J. M. Arias, C. Ratti, Predicting regional economic indices using big data of individual bank card transactions, in: 2017 IEEE International Conference on Big Data, IEEE Press, 2017, pp. 1313–1318.
- [462] I. T. Jolliffe, Principal Component Analysis, 2nd Edition, Springer, New York, NY, USA, 2002.
- [463] R. D. Clemente, M. Luengooroz, M. Travizano, S. Xu, B. Vaitla, M. C. González, Sequences of purchases in credit card data reveal lifestyles in urban populations, *Nature Communications* 9 (2018) 3330.
- [464] A. C. Iversen, P. Kraft, Does socio-economic status and health consciousness influence how women respond to health related messages in media?, *Health Education Research* 21 (5) (2006) 601–610.
- [465] S. C. Wangberg, H. K. Andreassen, H. U. Prokosch, S. M. Santana, T. Sørensen, C. E. Chronaki, Relations between Internet use, socio-economic status (SES), social support and subjective health, *Health Promotion International* 23 (1) (2008) 70–77.
- [466] R. M. Filho, G. R. Borges, J. M. Almeida, G. L. Pappa, Inferring user social class in online social networks, in: Proceedings of the 8th Workshop on Social Network Mining and Analysis, SNAKDD'14, ACM Press, New York, NY, USA, 2014, p. 10.
- [467] D. Preoțiu-Pietro, S. Volkova, V. Lampos, Y. Bachrach, N. Aletras, Studying user income through language, behaviour and affect in social media, *PLoS ONE* 10 (9) (2015) e0138717.
- [468] V. Lampos, N. Aletras, J. K. Geyti, B. Zou, I. J. Cox, Inferring the socioeconomic status of social media users based on behaviour and language, in: Proceedings of the 38th European Conference on Information Retrieval, ECIR 2016, Springer, Cham, Switzerland, 2016, pp. 689–695.
- [469] P. Elias, M. Birch, SOC2010: Revision of the Standard Occupational Classification, *Economic and Labour Market Review* 4 (4) (2010) 48–55.
- [470] Y. Leo, E. Fleury, C. Sarraute, J. I. Alvarez-Hamelin, M. Karsai, Socioeconomic correlations in communication networks, in: Proceedings of the Fourth Conference on the Analysis of Mobile Phone Datasets, NetMob 2015, MIT Media Lab, Cambridge, MA, USA, 2015, pp. 1–2.
- [471] M. Fixman, A. Berenstein, J. Brea, M. Minnoni, M. Travizano, C. Sarraute, A bayesian approach to income inference in a communication network, in: 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM'16, IEEE Press, 2016, pp. 579–582.
- [472] S. Luo, F. Morone, C. Sarraute, M. Travizano, H. A. Makse, Inferring personal economic status from social network location, *Nature Communications* 8 (2017) 15227.
- [473] F. Morone, H. A. Makse, Influence maximization in complex networks through optimal percolation, *Nature* 524 (7563) (2015) 65–68.
- [474] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, H. A. Makse, Identification of influential spreaders in complex networks, *Nature Physics* 6 (11) (2010) 888–893.
- [475] E. Jahani, G. Saint-Jacques, P. Sundsøy, J. Bjelland, E. Moro, A. S. Pentland, Differential network effects on economic outcomes: A structural perspective, in: Proceedings of the 9th International Conference on Social Informatics, SocInfo 2017, Springer, Cham, Switzerland, 2017, pp. 41–50.
- [476] Q. Wang, J. Gao, T. Zhou, Z. Hu, H. Tian, Critical size of ego communication networks, *EPL (Europhysics Letters)* 114 (5) (2016) 58004.
- [477] W.-J. Xie, Y.-H. Yang, M.-X. Li, Z.-Q. Jiang, W.-X. Zhou, Individual position diversity in dependence socioeconomic networks increases economic output, *EPJ Data Science* 6 (2017) 10.
- [478] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, U. Alon, Network motifs: Simple building blocks of complex networks, *Science* 298 (5594) (2002) 824–827.
- [479] D. B. Stouffer, M. Sales-Pardo, M. I. Sirer, J. Bascompte, Evolutionary conservation of species' roles in food webs, *Science* 335 (6075) (2012) 1489–1492.
- [480] J. Macqueen, Some methods for classification and analysis of multivariate observations, in: Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, California, CA, USA, 1967, pp. 281–297.
- [481] H. Barbosa, M. Barthélémy, G. Ghoshal, C. R. James, M. Lenormand, T. Louail, R. Menezes, J. J. Ramasco, F. Simini, M. Tomasini, Human mobility: Models and applications, *Physics Reports* 734 (2018) 1–74.
- [482] C. Fan, Y. Liu, J. Huang, Z.-H. Rong, T. Zhou, Correlation between social proximity and mobility similarity, *Scientific Reports* 7 (2017) 11975.
- [483] A. Carlsson-Kanyama, A.-L. Linden, Travel patterns and environmental effects now and in the future: Implications of differences in energy consumption among socio-economic groups, *Ecological Economics* 30 (3) (1999) 405–417.

- [484] C. Propper, M. Damiani, G. Leckie, J. Dixon, Impact of patients' socioeconomic status on the distance travelled for hospital admission in the English National Health Service, *Journal of Health Services Research and Policy* 12 (3) (2007) 153–159.
- [485] L. Lotero, A. Cardillo, R. Hurtado, J. Gómez-Gardeñes, Several multiplexes in the same city: The role of socioeconomic differences in urban mobility, in: A. Garas (Ed.), *Interconnected Networks*, Springer, Cham, Switzerland, 2016, pp. 149–164.
- [486] S. Boccaletti, G. Bianconi, R. Criado, C. I. Del Genio, J. Gómez-Gardenes, M. Romance, I. Sendina-Nadal, Z. Wang, M. Zanin, The structure and dynamics of multilayer networks, *Physics Reports* 544 (1) (2014) 1–122.
- [487] L. Lotero, R. G. Hurtado, L. M. Floría, Rich do not rise early: Spatio-temporal patterns in the mobility networks of different socio-economic classes, *Royal Society Open Science* 3 (10) (2016) 150654.
- [488] G. Carra, I. Mulalic, M. Fosgerau, M. Barthélémy, Modelling the relation between income and commuting distance, *Journal of the Royal Society Interface* 13 (119) (2016) 20160306.
- [489] W. Alonso, *Location and Land Use*, Harvard University Press, New York, NY, USA, 2009.
- [490] V. Frias-Martinez, J. Virseda-Jerez, E. Frias-Martinez, On the relation between socio-economic status and physical mobility, *Information Technology for Development* 18 (2) (2012) 91–106.
- [491] V. Frias-Martinez, C. Soguero-Ruiz, E. Frias-Martinez, M. Josephidou, Forecasting socioeconomic trends with cell phone records, in: Proceedings of the 3rd ACM Symposium on Computing for Development, ACM DEV'13, ACM Press, New York, NY, USA, 2013, p. 15.
- [492] L. Pappalardo, D. Pedreschi, Z. Smoreda, F. Giannotti, Using big data to study the link between human mobility and socio-economic development, in: 2015 IEEE International Conference on Big Data, IEEE Press, 2015, pp. 871–878.
- [493] C. Song, Z. Qu, N. Blumm, A.-L. Barabási, Limits of predictability in human mobility, *Science* 327 (5968) (2010) 1018–1021.
- [494] L. Pappalardo, M. Vanhoof, L. Gabrielli, Z. Smoreda, D. Pedreschi, F. Giannotti, An analytical framework to nowcast well-being using mobile phone data, *International Journal of Data Science and Analytics* 2 (1–2) (2016) 75–92.
- [495] M. Florez, S. Jiang, R. Li, C. H. Mojica, S. A. Transmilenio, R. A. Rios, M. C. González, Measuring the impact of economic well being in commuting networks—A case study of Bogota, Colombia, in: *Proceedings of the Transportation Research Board 96th Annual Meeting*, 2018, pp. 1–19.
- [496] X. Yang, A. Belyi, I. Bojic, C. Ratti, Human mobility and socioeconomic status: Analysis of Singapore and Boston, *Computers, Environment and Urban Systems* 72 (2018) 51–67.
- [497] L. Hong, E. Frias-Martinez, V. Frias-Martinez, Topic models to infer socio-economic maps, in: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI'16*, AAAI Press, Palo Alto, CA, USA, 2016, pp. 3835–3841.
- [498] V. K. Singh, B. Bozkaya, A. Pentland, Money walks: Implicit mobility behavior and financial well-being, *PLoS ONE* 10 (8) (2015) e0136628.
- [499] M. Lenormand, T. Louail, O. G. Cantú-Ros, M. Picornell, R. Herranz, J. M. Arias, M. Barthélémy, M. S. Miguel, J. J. Ramasco, Influence of sociodemographic characteristics on human mobility, *Scientific Reports* 5 (2015) 10075.
- [500] Y. Zhu, F. Chen, M. Li, Z. Wang, Inferring the economic attributes of urban rail transit passengers based on individual mobility using multisource data, *Sustainability* 10 (11) (2018) 4178.
- [501] M. G. Shahnawaz, H. Jafri, Job attitudes as predictor of employee turnover among stayers and leavers/hoppers, *Journal of Management Research* 9 (3) (2009) 159–166.
- [502] M. C. Sturman, L. Shao, J. H. Katz, The effect of culture on the curvilinear relationship between performance and turnover, *Journal of Applied Psychology* 97 (1) (2012) 46–62.
- [503] A.-L. Barabási, *The Formula: The Universal Laws of Success*, Hachette Book Group, New York, NY, USA, 2018.
- [504] M. Ettredge, J. Gerdes, G. Karuga, Using web-based search data to predict macroeconomic statistics, *Communications of the ACM* 48 (11) (2005) 87–92.
- [505] N. Askitas, K. F. Zimmermann, Google econometrics and unemployment forecasting, *Applied Economics Quarterly* 55 (2) (2009) 107–120.
- [506] H. Choi, H. Varian, Predicting initial claims for unemployment benefits, Tech. rep., Google Inc., Mountain View, CA, USA (2009).
- [507] H. Choi, H. Varian, Predicting the present with Google Trends, *Economic Record* 88 (s1) (2012) 2–9.
- [508] F. D'Amuri, Predicting unemployment in short samples with Internet job search query data, MPRA Paper No. 18403, University Library of Munich, Munich, Germany (2009).
- [509] F. D'Amuri, J. Marcucci, The predictive power of Google searches in forecasting US unemployment, *International Journal of Forecasting* 33 (4) (2017) 801–816.
- [510] W. Xu, Z. Li, Q. Chen, Forecasting the unemployment rate by neural networks using search engine query data, in: *2012 45th Hawaii International Conference on System Sciences*, IEEE Press, 2012, pp. 3591–3599.
- [511] W. Xu, Z. Li, C. Cheng, T. Zheng, Data mining for unemployment rate prediction using search engine query data, *Service Oriented Computing and Applications* 7 (1) (2013) 33–42.
- [512] N. Barreira, P. Godinho, P. Melo, Nowcasting unemployment rate and new car sales in south-western Europe with Google Trends, *NETNOMICS: Economic Research and Electronic Networking* 14 (3) (2014) 129–165.
- [513] Z. Li, W. Xu, L. Zhang, R. Y. K. Lau, An ontology-based web mining method for unemployment rate prediction, *Decision Support Systems* 66 (2014) 114–122.
- [514] A. L. Montgomery, V. Zarnowitz, R. S. Tsay, G. C. Tiao, Forecasting the u.s. unemployment rate, *Journal of the American Statistical Association* 93 (442) (1998) 478–493.
- [515] M. R. Vicente, A. J. López-Menéndez, R. Pérez, Forecasting unemployment with Internet search data: Does it help to improve predictions when job destruction is skyrocketing?, *Technological Forecasting and Social Change* 92 (2015) 132–139.
- [516] J. Pavlicek, L. Kristoufek, Nowcasting unemployment rates with Google searches: Evidence from the Visegrad Group countries, *PLoS ONE* 10 (5) (2015) e0127084.
- [517] S. Falorsi, A. Naccarato, A. Pierini, Using Google Trend data to predict the Italian unemployment rate, *Departmental Working Papers of Economics No. 203*, Department of Economics, University Roma Tre, Rome, Italy (2015).
- [518] M. G. Chadwick, G. Sengül, Nowcasting the unemployment rate in Turkey: Let's ask Google, *Central Bank Review* 15 (3) (2015) 15–40.
- [519] Y. Fondeur, F. Karamé, Can Google data help predict French youth unemployment?, *Economic Modelling* 30 (1) (2013) 117–125.

- [520] J. Durbin, S. J. Koopman, *Time Series Analysis by State Space Methods*, Oxford University Press, New York, NY, USA, 2001.
- [521] C.-M. Kwon, J. U. Jung, Forecasting youth unemployment in Korea with web search queries, in: Proceedings of the the 9th Conference on Internet of Things, Smart Spaces, and Next Generation Networks and Systems, ruSMART 2016, Springer, Cham, Switzerland, 2016, pp. 3–14.
- [522] A. Naccarato, S. Falorsi, S. Loriga, A. Pierini, Combining official and Google Trends data to forecast the Italian youth unemployment rate, *Technological Forecasting and Social Change* 130 (2018) 114–122.
- [523] H. Lütkepohl, *New Introduction to Multiple Time Series Analysis*, 2nd Edition, Springer, Heidelberg, Germany, 2006.
- [524] S. R. Baker, A. Fradkin, The impact of unemployment insurance on job search: Evidence from Google search data, *Review of Economics and Statistics* 99 (5) (2017) 756–768.
- [525] O. A. Guerrero, E. Lopez, Understanding unemployment in the era of big data: Policy informed by data-driven theory, *Policy and Internet* 9 (1) (2017) 28–54.
- [526] D. Antenucci, M. Cafarella, M. Levenstein, C. Ré, M. D. Shapiro, Using social media to measure labor market flows, Working Paper No. 20010, National Bureau of Economic Research, Cambridge, MA, USA (2014).
- [527] D. Proserpio, S. Counts, A. Jain, The psychology of job loss: Using social media data to characterize and predict unemployment, in: Proceedings of the 8th ACM Conference on Web Science, WebSci’16, ACM Press, New York, NY, USA, 2016, pp. 223–232.
- [528] A. Llorente, M. García-Herranz, M. Cebrian, E. Moro, Social media fingerprints of unemployment, *PLoS ONE* 10 (5) (2015) e0128692.
- [529] E. Bokányi, Z. Lábszki, G. Vattay, Prediction of employment and unemployment rates from Twitter daily rhythms in the US, *EPJ Data Science* 6 (2017) 14.
- [530] D. Krackhardt, L. W. Porter, The snowball effect: Turnover embedded in communication networks, *Journal of Applied Psychology* 71 (1) (1986) 50–55.
- [531] T. H. Feeley, G. A. Barnett, Predicting employee turnover from communication networks, *Human Communication Research* 23 (3) (1997) 370–387.
- [532] K. W. Mossholder, R. P. Settoon, S. C. Henagan, A relational perspective on turnover: Examining structural, attitudinal, and behavioral predictors, *Academy of Management Journal* 48 (4) (2005) 607–618.
- [533] T. H. Feeley, Testing a communication network model of employee turnover based on centrality, *Journal of Applied Communication Research* 28 (3) (2000) 262–277.
- [534] T. H. Feeley, S. I. Moon, R. S. Kozey, A. S. Slowe, An erosion model of employee turnover based on network centrality, *Journal of Applied Communication Research* 38 (2) (2010) 167–188.
- [535] J. Gao, L. Zhang, Q. M. Zhang, T. Zhou, Big data human resources: Performance analysis and promotion/resignation in employee networks, in: Y. Liu (Ed.), *Social Physics: Social Governance*, Science Press, Beijing, China, 2014, Ch. 4, pp. 38–56.
- [536] S. N. Dorogovtsev, A. V. Goltsev, J. F. F. Mendes,  $k$ -core organization of complex networks, *Physical Review Letters* 96 (2006) 040601.
- [537] L. Lü, Y.-C. Zhang, C. H. Yeung, T. Zhou, Leaders in social networks, the Delicious case, *PLoS ONE* 6 (6) (2011) e21202.
- [538] J. Yuan, Q.-M. Zhang, J. Gao, L. Zhang, X.-S. Wan, X.-J. Yu, T. Zhou, Promotion and resignation in employee networks, *Physica A: Statistical Mechanics and its Applications* 444 (2016) 442–447.
- [539] J. L. Toolet, Y.-R. Lin, E. Muehlegger, D. Shoag, M. C. González, D. Lazer, Tracking employment shocks using mobile phone data, *Journal of the Royal Society Interface* 12 (107) (2015) 20150185.
- [540] P. Sundsøy, J. Bjelland, B.-A. Reme, E. Jahani, E. Wetter, L. Bengtsson, Towards real-time prediction of unemployment and profession, in: Proceedings of the 9th International Conference on Social Informatics, SocInfo 2017, Springer, Cham, Switzerland, 2017, pp. 14–23.
- [541] A. Almatouq, F. Prieto-Castrillo, A. S. Pentland, Mobile communication signatures of unemployment, in: Proceedings of the 8th International Conference on Social Informatics, SocInfo 2016, Springer, Cham, Switzerland, 2016, pp. 407–418.
- [542] B. Yucesoy, A.-L. Barabási, Untangling performance from success, *EPJ Data Science* 5 (2016) 17.
- [543] R. T. Sparrowe, R. C. Liden, S. J. Wayne, Social networks and the performance of individuals and groups, *Academy of Management Journal* 44 (2) (2001) 316–325.
- [544] M. K. Ahuja, D. F. Galletta, K. M. Carley, Individual centrality and performance in virtual R&D groups: An empirical study, *Management Science* 49 (1) (2003) 21–38.
- [545] B. L. Kirkman, B. Rosen, P. E. Tesluk, C. B. Gibson, The impact of team empowerment on virtual team performance: The moderating role of face-to-face interaction, *Academy of Management Journal* 47 (2) (2004) 175–192.
- [546] R. Cross, J. N. Cummings, Tie and network correlates of individual performance in knowledge-intensive work, *Academy of Management Journal* 47 (6) (2004) 928–937.
- [547] J. Duch, J. S. Waitzman, L. A. N. Amaral, Quantifying the performance of individual players in a team activity, *PLoS ONE* 5 (6) (2010) e10937.
- [548] A. W. Woolley, C. F. Chabris, A. Pentland, N. Hashmi, T. W. Malone, Evidence for a collective intelligence factor in the performance of human groups, *Science* 330 (6004) (2010) 686–688.
- [549] J. B. Bear, A. W. Woolley, The role of gender in team collaboration and performance, *Interdisciplinary Science Reviews* 36 (2) (2011) 146–153.
- [550] M. Cai, H. Du, C. Zhao, W. Du, Relationship between employees' performance and social network structure: An empirical research based on a SME from a whole-network perspective, *Chinese Management Studies* 8 (1) (2014) 85–108.
- [551] M. Cai, W. Wang, Y. Cui, H. E. Stanley, Multiplex network analysis of employee performance and employee social relationships, *Physica A: Statistical Mechanics and its Applications* 490 (2018) 1–12.
- [552] A. Mao, W. Mason, S. Suri, D. J. Watts, An experimental study of team size and performance on a complex task, *PLoS ONE* 11 (4) (2016) e0153048.
- [553] A. S. Pentland, The new science of building great teams, *Harvard Business Review* 90 (4) (2012) 60–69.
- [554] D. O. Olguín, B. N. Waber, T. Kim, A. Mohan, K. Ara, A. Pentland, Sensible organizations: Technology and methodology for automatically measuring organizational behavior, *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 39 (1) (2009) 43–55.
- [555] J.-i. Watanabe, M. Fujita, K. Yano, H. Kanesaka, T. Hasegawa, Resting time activeness determines team performance in call centers, in:

- 2012 International Conference on Social Informatics, IEEE Press, 2012, pp. 26–31.
- [556] D. Tjosvold, N. Y. Chen, X. Huang, D. Xu, Developing cooperative teams to support individual performance and well-being in a call center in China, *Group Decision and Negotiation* 23 (2) (2014) 325–348.
- [557] Y.-A. de Montjoye, A. Stopczynski, E. Shmueli, A. Pentland, S. Lehmann, The strength of the strongest ties in collaborative problem solving, *Scientific Reports* 4 (2014) 5277.
- [558] A. Zeng, Z. Shen, J. Zhou, J. Wu, Y. Fan, Y. Wang, H. E. Stanley, The science of science: From the perspective of complex systems, *Physics Reports* 714–715 (2017) 1–73.
- [559] S. Fortunato, C. T. Bergstrom, K. Börner, J. A. Evans, D. Helbing, S. Milojević, A. M. Petersen, F. Radicchi, R. Sinatra, B. Uzzi, A. Vespignani, L. Waltman, D. Wang, A.-L. Barabási, Science of science, *Science* 359 (6379) (2018) eaao0185.
- [560] S. Wuchty, B. Jones, B. Uzzi, The increasing dominance of teams in production of knowledge, *Science* 316 (5827) (2007) 1036–1039.
- [561] B. F. Jones, S. Wuchty, B. Uzzi, Multi-university research teams: Shifting impact, geography, and stratification in science, *Science* 322 (5905) (2008) 1259–1262.
- [562] D. de Stefano, V. Fuccella, M. P. Vitale, S. Zaccarin, The use of different data sources in the analysis of co-authorship networks and scientific performance, *Social Networks* 35 (3) (2013) 370–381.
- [563] A. Lunguau, Y. Huang, N. S. Contractor, Understanding the assembly of interdisciplinary teams and its impact on performance, *Journal of Informetrics* 8 (1) (2014) 59–70.
- [564] J. E. Hirsch, An index to quantify an individual's scientific research output, *Proceedings of the National Academy of Sciences of the United States of America* 102 (46) (2005) 16569–16572.
- [565] J. E. Hirsch, Does the h index have predictive power?, *Proceedings of the National Academy of Sciences of the United States of America* 104 (49) (2007) 19193–19198.
- [566] F. Radicchi, S. Fortunato, C. Castellano, Universality of citation distributions: Toward an objective measure of scientific impact, *Proceedings of the National Academy of Sciences of the United States of America* 105 (45) (2008) 17268–17272.
- [567] A. Abbasi, J. Altmann, J. Hwang, Evaluating scholars based on their academic collaboration activities: Two indices, the RC-index and the CC-index, for quantifying collaboration activities of researchers and scientific communities, *Scientometrics* 83 (1) (2010) 1–13.
- [568] L. Bornmann, R. Mutz, S. E. Hug, H.-D. Daniel, A multilevel meta-analysis of studies reporting correlations between the h index and 37 different h index variants, *Journal of Informetrics* 5 (3) (2011) 346–359.
- [569] D. Wang, C. Song, A.-L. Barabási, Quantifying long-term scientific impact, *Science* 342 (6154) (2013) 127–132.
- [570] R. Sinatra, D. Wang, P. Deville, C. Song, A.-L. Barabási, Quantifying the evolution of individual scientific impact, *Science* 354 (6312) (2016) aaf5239.
- [571] P. Deville, D. Wang, R. Sinatra, C. Song, V. D. Blondel, A.-L. Barabási, Career on the move: Geography, stratification, and scientific impact, *Scientific Reports* 4 (2014) 4770.
- [572] H.-W. Shen, A.-L. Barabási, Collective credit allocation in science, *Proceedings of the National Academy of Sciences of the United States of America* 111 (34) (2014) 12325–12330.
- [573] T. Jia, D. Wang, B. Szymanski, Quantifying patterns of research-interest evolution, *Nature Human Behaviour* 1 (2017) 78.
- [574] R. Wang, G. Harari, P. Hao, X. Zhou, A. T. Campbell, SmartGPA: How smartphones can assess and predict academic performance of college students, in: *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp'15*, ACM Press, New York, NY, USA, 2015, pp. 295–306.
- [575] Y. Cao, J. Gao, D. Lian, Z. Rong, J. Shi, Q. Wang, Y. Wu, H. Yao, T. Zhou, Orderliness predicts academic performance: Behavioural analysis on campus lifestyle, *Journal of the Royal Society Interface* 15 (146) (2018) 20180210.
- [576] Y. Cao, J. Gao, T. Zhou, Orderliness of campus lifestyle predicts academic performance: A case study in Chinese university, in: H. Baumeister, C. Montag (Eds.), *Mobile Sensing and Digital Phenotyping: New Development in Psychoinformatics*, Springer Nature Switzerland AG, Basel, Switzerland, 2019, p. In press.
- [577] I. Kontoyiannis, P. H. Algoet, Y. M. Suhov, A. J. Wyner, Nonparametric entropy estimation for stationary processes and random fields, with applications to English text, *IEEE Transactions on Information Theory* 44 (3) (1998) 1319–1327.
- [578] P. Xu, L. Yin, Z. Yue, T. Zhou, On predictability of time series, *Physica A: Statistical Mechanics and its Applications* 523 (2019) 345–351.
- [579] H. Yao, D. Lian, Y. Cao, Y. Wu, T. Zhou, Predicting academic performance for college students: A campus behavior perspective, *ACM Transactions on Intelligent Systems and Technology* 10 (3) (2019) 24.
- [580] D. Rao, D. Yarowsky, A. Shreevats, M. Gupta, Classifying latent user attributes in Twitter, in: *Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents, SMUC'10*, ACM Press, New York, NY, USA, 2010, pp. 37–44.
- [581] N. Garera, D. Yarowsky, Modeling latent biographic attributes in conversational genres, in: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL, ACL'09*, Association for Computational Linguistics, Stroudsburg, PA, USA, 2009, pp. 710–718.
- [582] J. D. Burger, J. Henderson, G. Kim, G. Zarrella, Discriminating gender on Twitter, in: *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP'11*, Association for Computational Linguistics, Stroudsburg, PA, USA, 2011, pp. 1301–1309.
- [583] W. Liu, D. Ruths, What's in a name? Using first names as features for gender inference in Twitter, in: *Proceedings of the 2013 AAAI Spring Symposium: Analyzing Microtext, AAAI'13*, AAAI Press, Palo Alto, CA, USA, 2013, pp. 10–16.
- [584] M. Ciot, M. Sonderegger, D. Ruths, Gender inference of Twitter users in non-English contexts, in: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP'13*, Association for Computational Linguistics, Stroudsburg, PA, USA, 2013, pp. 1136–1145.
- [585] S. Volkova, Y. Bachrach, M. Armstrong, V. Sharma, Inferring latent user properties from texts published in social media, in: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, AAAI'15*, AAAI Press, Palo Alto, CA, USA, 2015, pp. 4296–4297.
- [586] A. Culotta, N. K. Ravi, J. Cutler, Predicting the demographics of Twitter users from website traffic data, in: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, AAAI'15*, AAAI Press, Palo Alto, CA, USA, 2015, pp. 72–78.
- [587] O. Montasser, D. Kifer, Predicting demographics of high-resolution geographies with geotagged tweets, in: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI'17*, AAAI Press, Palo Alto, CA, USA, 2017, pp. 1460–1466.
- [588] Y. Huang, L. Yu, X. Wang, B. Cui, A multi-source integration framework for user occupation inference in social media systems, *World Wide Web* 22 (2018) 103–125.

- Web 18 (5) (2015) 1247–1267.
- [589] D. Preotiuc-Pietro, V. Lampis, N. Aletras, An analysis of the user occupational class through Twitter content, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Stroudsburg, PA, USA, 2015, pp. 1754–1764.
- [590] L. Sloan, J. Morgan, P. Burnap, M. Williams, Who tweets? Deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data, *PLoS ONE* 10 (3) (2015) e0115545.
- [591] V. Frias-Martinez, E. Frias-Martinez, N. Oliver, A gender-centric analysis of calling behavior in a developing economy using call detail records, in: Proceedings of the 2010 AAAI Spring Symposium: Artificial Intelligence for Development, AID’10, AAAI Press, Palo Alto, CA, USA, 2010, pp. 37–42.
- [592] C. Herrera-Yagüe, P. J. Zufiria, Prediction of telephone user attributes based on network neighborhood information, in: Proceedings of the 8th International Conference on Machine Learning and Data Mining in Pattern Recognition, MLMR 2012, Springer, Berlin, Germany, 2012, pp. 645–659.
- [593] Y. Dong, Y. Yang, J. Tang, Y. Yang, N. V. Chawla, Inferring user demographics and social strategies in mobile social networks, in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD’14, ACM, New York, NY, USA, 2014, pp. 15–24.
- [594] Y. Dong, N. V. Chawla, J. Tang, Y. Yang, User modeling on demographic attributes in big mobile social networks, *ACM Transactions on Information Systems* 35 (4) (2017) 35.
- [595] C. Sarraute, P. Blanc, J. Burroni, A study of age and gender seen through mobile phone usage patterns in Mexico, in: 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, IEEE Press, 2014, pp. 836–843.
- [596] E. Jahani, P. Sundsøy, J. Bjelland, L. Bengtsson, A. S. Pentland, Y.-A. de Montjoye, Improving official statistics in emerging markets using machine learning and mobile phone data, *EPJ Data Science* 6 (2017) 3.
- [597] S. Akter, L. Holder, Using graphical features to improve demographic prediction from smart phone data, in: Proceedings of the 2nd International Workshop on Network Data Analytics, NDA’17, ACM Press, New York, NY, USA, 2017, p. 5.
- [598] P. Wang, F. Sun, D. Wang, J. Tao, X. Guan, A. Bifet, Inferring demographics and social networks of mobile device users on campus from AP-trajectories, in: Proceedings of the 26th International Conference on World Wide Web Companion, WWW’17 Companion, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 2017, pp. 139–147.
- [599] B. Felbo, P. Sundsøy, A. S. Pentland, S. Lehmann, Y.-A. de Montjoye, Modeling the temporal nature of human behavior for demographics prediction, in: Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases, ECML PKDD 2017, Springer, Cham, Switzerland, 2017, pp. 140–152.
- [600] Y. Zhong, N. J. Yuan, W. Zhong, F. Zhang, X. Xie, You are where you go: Inferring demographic attributes from location check-ins, in: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, WSDM’15, ACM Press, New York, NY, USA, 2015, pp. 295–304.
- [601] C. W. Lin, C. W. Lin, R. Nkambou, Inferring social network user profiles using a partial social graph, *Journal of Intelligent Information Systems* 47 (2) (2016) 313–344.
- [602] Y. Ren, M. Tomko, F. D Salim, J. Chan, M. Sanderson, Understanding the predictability of user demographics from cyber-physical-social behaviours in indoor retail spaces, *EPJ Data Science* 7 (2018) 1.
- [603] J. Hinds, A. N. Joinson, What demographic attributes do our digital footprints reveal? A systematic review, *PLoS ONE* 13 (11) (2018) e0207112.
- [604] R. R. McCrae, O. P. John, An introduction to the Five-Factor Model and its applications, *Journal of Personality* 60 (2) (1992) 175–215.
- [605] J. M. Digman, Personality structure: Emergence of the five-factor model, *Psychology* 41 (1) (2003) 417–440.
- [606] C. Ross, E. S. Orr, M. Susic, J. M. Arseneault, M. G. Simmering, R. R. Orr, Personality and motivations associated with Facebook use, *Computers in Human Behavior* 25 (2) (2009) 578–586.
- [607] T. Correa, A. W. Hinsley, H. G. de Zuniga, Who interacts on the Web?: The intersection of users’ personality and social media use, *Computers in Human Behavior* 26 (2) (2010) 247–253.
- [608] J. Golbeck, C. Robles, K. Turner, Predicting personality with social media, in: CHI’11 Extended Abstracts on Human Factors in Computing Systems, CHI EA’11, ACM Press, New York, NY, USA, 2011, pp. 253–262.
- [609] Y. Bachrach, M. Kosinski, T. Graepel, P. Kohli, D. Stillwell, Personality and patterns of Facebook usage, in: Proceedings of the 4th Annual ACM Web Science Conference, WebSci’12, ACM Press, New York, NY, USA, 2012, pp. 24–32.
- [610] G. Seidman, Self-presentation and belonging on Facebook: How personality influences social media use and motivations, *Personality and Individual Differences* 54 (3) (2013) 402–407.
- [611] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. P. Seligman, L. H. Ungar, Personality, gender, and age in the language of social media: The open-vocabulary approach, *PLoS ONE* 8 (9) (2013) e73791.
- [612] A. E. Hoerl, R. W. Kennard, Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics* 42 (1) (2000) 80–86.
- [613] D. Preotiuc-Pietro, J. Carpenter, S. Giorgi, L. Ungar, Studying the dark triad of personality through Twitter behavior, in: Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM’16, ACM Press, New York, NY, USA, 2016, pp. 761–770.
- [614] D. L. Paulhus, K. M. Williams, The dark triad of personality: Narcissism, Machiavellianism, and psychopathy, *Journal of Research in Personality* 36 (6) (2002) 556–563.
- [615] D. Garcia, S. Sikström, The dark side of Facebook: Semantic representations of status updates predict the dark triad of personality, *Personality and Individual Differences* 67 (2014) 92–96.
- [616] S. C. Guntuku, W. Lin, J. Carpenter, W. K. Ng, L. H. Ungar, D. Preotiuc-Pietro, Studying personality through the content of posted and liked images on Twitter, in: Proceedings of the 2017 ACM on Web Science Conference, WebSci’17, ACM Press, New York, NY, USA, 2017, pp. 223–227.
- [617] L. Liu, D. Preotiuc-Pietro, Z. R. Samani, M. E. Moghaddam, L. H. Ungar, Analyzing personality through social media profile picture choice,

- in: Proceedings of the Tenth International AAAI Conference on Web and Social Media, ICWSM'16, AAAI Press, Palo Alto, CA, USA, 2016, pp. 211–220.
- [618] E. Zhou, H. Fan, Z. Cao, Y. Jiang, Q. Yin, Extensive facial landmark localization with coarse-to-fine convolutional network cascade, in: Proceedings of the 2013 IEEE International Conference on Computer Vision Workshops, ICCVW'13, IEEE Computer Society, Washington, D.C., USA, 2013, pp. 386–391.
- [619] C. Segalin, F. Celli, L. Polonio, M. Kosinski, D. Stillwell, N. Sebe, M. Cristani, B. Lepri, What your Facebook profile picture reveals about your personality, in: Proceedings of the 2017 ACM on Multimedia Conference, MM'17, ACM Press, New York, NY, USA, 2017, pp. 460–468.
- [620] H. Masum, Y.-C. Zhang, Manifesto for the reputation society, *First Monday* 9 (7) (2004) 1158.
- [621] L. Mui, M. Mohtashemi, A. Halberstadt, Notions of reputation in multi-agents systems: A review, in: Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS '02, ACM Press, New York, NY, USA, 2002, pp. 280–287.
- [622] G. Zacharia, A. Moukas, P. Maes, Collaborative reputation mechanisms for electronic marketplaces, *Decision Support Systems* 29 (4) (2000) 371–388.
- [623] J. Sabater, C. Sierra, Review on computational trust and reputation models, *Artificial Intelligence Review* 24 (1) (2005) 33–60.
- [624] P. Resnick, K. Kuwabara, R. Zeckhauser, E. Friedman, Reputation systems, *Communications of the ACM* 43 (12) (2000) 45–48.
- [625] A. JøSang, R. Ismail, C. Boyd, A survey of trust and reputation systems for online service provision, *Decision Support Systems* 43 (2) (2007) 618–644.
- [626] P. Laureti, L. Moret, Y.-C. Zhang, Y.-K. Yu, Information filtering via iterative refinement, *EPL (Europhysics Letters)* 75 (6) (2006) 1006.
- [627] H. Liao, A. Zeng, R. Xiao, Z.-M. Ren, D.-B. Chen, Y.-C. Zhang, Ranking reputation and quality in online rating systems, *PLoS ONE* 9 (5) (2014) e97146.
- [628] X.-L. Liu, Q. Guo, L. Hou, C. Cheng, J.-G. Liu, Ranking online quality and reputation via the user activity, *Physica A: Statistical Mechanics and its Applications* 436 (2015) 629–636.
- [629] X.-L. Liu, J.-G. Liu, K. Yang, Q. Guo, J.-T. Han, Identifying online user reputation of user-object bipartite networks, *Physica A: Statistical Mechanics and its Applications* 467 (2017) 508–516.
- [630] Y. Tian, J. Zhu, Learning from crowds in the presence of schools of thought, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'12, ACM Press, New York, NY, USA, 2012, pp. 226–234.
- [631] J. Gao, Y.-W. Dong, M.-S. Shang, S.-M. Cai, T. Zhou, Group-based ranking method for online rating systems with spamming attacks, *EPL (Europhysics Letters)* 110 (2) (2015) 28003.
- [632] J. Gao, T. Zhou, Evaluating user reputation in online rating systems via an iterative group-based ranking method, *Physica A: Statistical Mechanics and its Applications* 473 (2017) 546–560.
- [633] L. Dai, Q. Guo, X.-L. Liu, J.-G. Liu, Y.-C. Zhang, Identifying online user reputation in terms of user preference, *Physica A: Statistical Mechanics and its Applications* 494 (2018) 403–409.
- [634] Y.-L. Zhang, Q. Guo, J. Ni, J.-G. Liu, Memory effect of the online rating for movies, *Physica A: Statistical Mechanics and its Applications* 417 (2015) 261–266.
- [635] F. Fouss, Y. Achbany, M. Saerens, A probabilistic reputation model based on transaction ratings, *Information Sciences* 180 (11) (2010) 2095–2123.
- [636] H. Liao, G. Cimini, M. Medo, Measuring quality, reputation and trust in online communities, in: Proceedings of the 20th International Conference on Foundations of Intelligent Systems, ISMIS'12, Springer, Berlin, Heidelberg, 2012, pp. 405–414.
- [637] B. Li, R.-H. Li, I. King, M. R. Lyu, J. X. Yu, A topic-biased user reputation model in rating systems, *Knowledge and Information Systems* 44 (3) (2015) 581–607.
- [638] R.-H. Li, J. Xu Yu, X. Huang, H. Cheng, Robust reputation-based ranking on bipartite rating networks, in: Proceedings of the 2012 SIAM International Conference on Data Mining, SDM'12, SIAM/Omnipress, Philadelphia, PA, USA, 2012, pp. 612–623.
- [639] M. Medo, J. R. Wakeling, The effect of discrete vs. continuous-valued ratings on reputation and ranking systems, *EPL (Europhysics Letters)* 91 (4) (2010) 48004.
- [640] H. Liao, M. S. Mariani, M. Medo, Y.-C. Zhang, M.-Y. Zhou, Ranking in evolving complex networks, *Physics Reports* 689 (2017) 1–54.
- [641] M. Thelwall, D. Wilkinson, S. Uppal, Data mining emotion in social network communication: Gender differences in MySpace, *Journal of the Association for Information Science and Technology* 61 (1) (2010) 190–199.
- [642] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, A. Kappas, Sentiment strength detection in short informal text, *Journal of the Association for Information Science and Technology* 61 (12) (2010) 2544–2558.
- [643] A. Pak, P. Paroubek, Twitter as a corpus for sentiment analysis and opinion mining, in: Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010, European Language Resources Association, Paris, France, 2010, pp. 1320–1326.
- [644] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, J. N. Rosenquist, Pulse of the nation: US mood throughout the day inferred from Twitter, Tech. rep., Northeastern University, Boston, MA, USA (2010).
- [645] J. Bollen, H. Mao, A. Pepe, Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena, in: Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, ICWSM'11, AAAI Press, Palo Alto, CA, USA, 2011, pp. 450–453.
- [646] W. Wang, L. Chen, K. Thirunarayan, A. P. Sheth, Harnessing Twitter “big data” for automatic emotion identification, in: 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing, IEEE Press, 2012, pp. 587–592.
- [647] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, C.-J. Lin, LIBLINEAR: A library for large linear classification, *Journal of Machine Learning Research* 9 (9) (2008) 1871–1874.
- [648] I. H. Witten, E. Frank, M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, 3rd Edition, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2011.
- [649] M. E. Larsen, T. W. Boonstra, P. J. Batterham, B. O'Dea, C. Paris, H. Christensen, We feel: Mapping emotion on Twitter, *IEEE Journal of Biomedical and Health Informatics* 19 (4) (2015) 1246–1252.
- [650] N. M. Jones, S. P. Wojcik, J. Sweeting, R. C. Silver, Tweeting negative emotion: An investigation of Twitter data in the aftermath of violence on college campuses, *Psychological Methods* 21 (4) (2016) 526–541.

- [651] S. M. Mohammad, F. Bravomarquez, Emotion intensities in tweets, in: Proceedings of the 6th Joint Conference on Lexical and Computational Semantics, Association for Computational Linguistics, Stroudsburg, PA, USA, 2017, pp. 65–77.
- [652] S. Kiritchenko, S. Mohammad, Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Stroudsburg, PA, USA, 2017, pp. 465–470.
- [653] S. Madisetti, M. S. Desarkar, An ensemble based method for predicting emotion intensity of tweets, in: Proceedings of the 5th International Conference on Mining Intelligence and Knowledge Exploration, MIKE 2017, Springer, Cham, Switzerland, 2017, pp. 359–370.
- [654] Y. Kim, Convolutional neural networks for sentence classification, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Stroudsburg, PA, USA, 2014, pp. 1746–1751.
- [655] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD’16, ACM Press, New York, NY, USA, 2016, pp. 785–794.
- [656] M. Settanni, D. Marengo, Sharing feelings online: Studying emotional well-being via automated text analysis of Facebook posts, *Frontiers in Psychology* 6 (2015) 1045.
- [657] S. Zheng, J. Wang, C. Sun, X. Zhang, M. E. Kahn, Air pollution lowers Chinese urbanites' expressed happiness on social media, *Nature Human Behaviour* 3 (2019) 237–243.
- [658] G. Coppersmith, M. Dredze, C. Harman, Quantifying mental health signals in Twitter, in: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, Association for Computational Linguistics, Stroudsburg, PA, USA, 2014, pp. 51–60.
- [659] J. W. Pennebaker, C. K. Chung, M. Ireland, A. Gonzales, R. J. Booth, The development and psychometric properties of LIWC2007, Tech. rep., LIWC.net, Austin, TX, USA (2007).
- [660] G. Coppersmith, M. Dredze, C. Harman, K. Hollingshead, From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses, in: Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, Association for Computational Linguistics, Stroudsburg, PA, USA, 2015, pp. 1–10.
- [661] S. Balani, M. de Choudhury, Detecting and characterizing mental health related self-disclosure in social media, in: Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems, CHI EA’15, ACM Press, New York, NY, USA, 2015, pp. 1373–1378.
- [662] M. A. Moreno, L. A. Jelenchick, K. G. Egan, E. Cox, H. Young, K. E. Gannon, T. Becker, Feeling bad on Facebook: Depression disclosures by college students on a social networking site, *Depression and Anxiety* 28 (6) (2011) 447–455.
- [663] S. Park, S. W. Lee, J. Kwak, M. Cha, B. Jeong, Activities on Facebook reveal the depressive state of users, *Journal of Medical Internet Research* 15 (10) (2013) e217.
- [664] L. S. Radloff, The CES-D scale: A self-report depression scale for research in the general population, *Applied Psychological Measurement* 1 (3) (1977) 385–401.
- [665] M. de Choudhury, S. Counts, E. J. Horvitz, A. Hoff, Characterizing and predicting postpartum depression from shared Facebook data, in: Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW’14, ACM Press, New York, NY, USA, 2014, pp. 626–638.
- [666] M. de Choudhury, M. Gamon, S. Counts, E. Horvitz, Predicting depression via social media, in: Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media, ICWSM’13, AAAI Press, Palo Alto, CA, USA, 2013, pp. 128–137.
- [667] R. O. Duda, P. E. Hart, D. G. Stork, *Pattern Classification*, 2nd Edition, Wiley-Interscience, New York, NY, USA, 2000.
- [668] X. Wang, C. Zhang, Y. Ji, L. Sun, L. Wu, Z. Bao, A depression detection model based on sentiment analysis in micro-blog social network, in: Proceedings of the PAKDD 2013 International Workshops, PAKDD 2013, Springer, Berlin, Heidelberg, 2013, pp. 201–213.
- [669] S. Tsugawa, Y. Kikuchi, F. Kishino, K. Nakajima, Y. Itoh, H. Ohsaki, Recognizing depression from Twitter activity, in: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI’15, ACM Press, New York, NY, USA, 2015, pp. 3187–3196.
- [670] A. G. Reece, C. M. Danforth, Instagram photos reveal predictive markers of depression, *EPJ Data Science* 6 (2017) 15.
- [671] A. G. Reece, A. J. Reagan, K. L. Lix, P. S. Dodds, C. M. Danforth, E. J. Langer, Forecasting the onset and course of mental illness with Twitter data, *Scientific Reports* 7 (2017) 13006.
- [672] H.-H. Won, W. Myung, G.-Y. Song, W.-H. Lee, J.-W. Kim, B. J. Carroll, D. K. Kim, Predicting national suicide numbers with social media data, *PLoS ONE* 8 (4) (2013) e61809.
- [673] H. Sueki, The association of suicide-related Twitter use with suicidal behaviour: A cross-sectional study of young Internet users in Japan, *Journal of Affective Disorders* 170 (2015) 155–160.
- [674] A. Abboute, Y. Boudjeriou, G. Entringer, J. Azé, S. Bringay, P. Poncelet, Mining Twitter for suicide prevention, in: Proceedings of the 19th International Conference on Applications of Natural Language to Information Systems, NLDB 2014, Springer, Cham, Switzerland, 2014, pp. 250–253.
- [675] J. F. Gunn, D. Lester, Twitter postings and suicide: An analysis of the postings of a fatal suicide in the 24 hours prior to death, *Suicidology* 17 (3) (2012) 28–30.
- [676] J. Jashinsky, S. H. Burton, C. L. Hanson, J. West, C. Giraud-Carrier, M. D. Barnes, T. Argyle, Tracking suicide risk factors through twitter in the US, *Crisis: The Journal of Crisis Intervention and Suicide Prevention* 35 (1) (2014) 51–59.
- [677] P. Burnap, W. Colombo, J. Scourfield, Machine classification and analysis of suicide-related communication on Twitter, in: Proceedings of the 26th ACM Conference on Hypertext and Social Media, HT’15, ACM Press, New York, NY, USA, 2015, pp. 75–84.
- [678] A. Benton, M. Mitchell, D. Hovy, Multitask learning for mental health conditions with limited social media data, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Vol. 1 of EACL’17, Association for Computational Linguistics, Stroudsburg, PA, USA, 2017, pp. 152–162.
- [679] L. Kristoufek, H. S. Moat, T. Preis, Estimating suicide occurrence statistics using Google Trends, *EPJ Data Science* 5 (2016) 32.
- [680] U. S. Tran, R. Andel, T. Niederkrotenthaler, B. Till, V. Ajdacic-Gross, M. Voracek, Low validity of Google Trends for behavioral forecasting of national suicide rates, *PLoS ONE* 12 (8) (2017) e0183149.
- [681] J. Robinson, G. Cox, E. Bailey, S. Hetrick, M. Rodrigues, S. Fisher, H. Herrman, Social media and suicide prevention: A systematic review,

- Early Intervention in Psychiatry 10 (2) (2016) 103–121.
- [682] D. C. Mohr, M. Zhang, S. M. Schueller, Personal sensing: Understanding mental health using ubiquitous sensors and machine learning, *Annual Review of Clinical Psychology* 13 (2017) 23–47.
- [683] R. Melia, K. Francis, J. Duggan, J. Bogue, M. O’Sullivan, D. Chambers, K. Young, Mobile health technology interventions for suicide prevention: Protocol for a systematic review and meta-analysis, *JMIR Research Protocols* 7 (1) (2018) e28.
- [684] K. E. Jones, N. G. Patel, M. A. Levy, A. Storeygard, D. Balk, J. L. Gittleman, P. Daszak, Global trends in emerging infectious diseases, *Nature* 451 (7181) (2008) 990–993.
- [685] D. Brockmann, D. Helbing, The hidden geometry of complex, network-driven contagion phenomena, *Science* 342 (6164) (2013) 1337–1342.
- [686] H.-F. Zhang, Z. Yang, Z.-X. Wu, B.-H. Wang, T. Zhou, Braess’s paradox in epidemic game: Better condition results in less payoff, *Scientific Reports* 3 (2013) 3292.
- [687] R. Pastor-Satorras, C. Castellano, P. van Mieghem, A. Vespignani, Epidemic processes in complex networks, *Reviews of Modern Physics* 87 (3) (2015) 925.
- [688] W. Wang, M. Tang, H. E. Stanley, L. A. Braunstein, Unification of theoretical approaches for epidemic spreading on complex networks, *Reports on Progress in Physics* 80 (3) (2017) 036603.
- [689] W. Wang, M. Cai, M. Zheng, Social contagions on correlated multiplex networks, *Physica A: Statistical Mechanics and its Applications* 499 (2018) 121–128.
- [690] W. Wang, Q.-H. Liu, J. Liang, Y. Hu, T. Zhou, Coevolution spreading in complex networks, *arXiv:1901.02125*, 2019.
- [691] S. Funk, M. Salathé, V. A. Jansen, Modelling the influence of human behaviour on the spread of infectious diseases: A review, *Journal of the Royal Society Interface* 7 (2010) 1247–1256.
- [692] V. Belik, T. Geisel, D. Brockmann, Natural human mobility patterns and spatial spread of infectious diseases, *Physical Review X* 1 (1) (2011) 011001.
- [693] P. C. Pinto, P. Thiran, M. Vetterli, Locating the source of diffusion in large-scale networks, *Physical Review Letters* 109 (6) (2012) 068702.
- [694] Z. Shen, S. Cao, W.-X. Wang, Z. Di, H. E. Stanley, Locating the source of diffusion in complex networks by time-reversal backward spreading, *Physical Review E* 93 (3) (2016) 032301.
- [695] J. Shaman, A. Karspeck, Forecasting seasonal outbreaks of influenza, *Proceedings of the National Academy of Sciences of the United States of America* 109 (50) (2012) 20425–20430.
- [696] H.-F. Zhang, J.-R. Xie, M. Tang, Y.-C. Lai, Suppression of epidemic spreading in complex networks by local information based behavioral responses, *Chaos: An Interdisciplinary Journal of Nonlinear Science* 24 (4) (2014) 043106.
- [697] Z. Wang, C. T. Bauch, S. Bhattacharyya, A. d’Onofrio, P. Manfredi, M. Perc, N. Perra, M. Salathé, D. Zhao, Statistical physics of vaccination, *Physics Reports* 664 (2016) 1–113.
- [698] X. Chen, W. Wei, S. Cai, H. E. Stanley, L. A. Braunstein, Optimal resource diffusion for suppressing disease spreading in multiplex networks, *Journal of Statistical Mechanics: Theory and Experiment* 2018 (5) (2018) 053501.
- [699] B. M. Althouse, S. V. Scarpino, L. A. Meyers, J. W. Ayers, M. Bargsten, J. Baumbach, J. S. Brownstein, L. Castro, H. Clapham, D. A. Cummings, S. Del Valle, S. Eubank, G. Fairchild, L. Finelli, N. Generous, D. George, D. R. Harper, L. Hébert-Dufresne, M. A. Johansson, K. Konty, M. Lipsitch, G. Milinovich, J. D. Miller, E. O. Nsoesie, D. R. Olson, M. Paul, P. M. Polgreen, R. Priedhorsky, J. M. Read, I. Rodríguez-Barraquer, D. J. Smith, C. Stefansen, D. L. Swerdlow, D. Thompson, A. Vespignani, A. Wesolowski, Enhancing disease surveillance with novel data streams: Challenges and opportunities, *EPJ Data Science* 4 (2015) 17.
- [700] E. C. Lee, A. Arab, S. M. Goldlust, C. Viboud, B. T. Grenfell, S. Bansal, Deploying digital health data to optimize influenza surveillance at national and local scales, *PLoS Computational Biology* 14 (3) (2018) e1006020.
- [701] M. Santillana, E. O. Nsoesie, S. R. Mekaru, D. Scales, J. S. Brownstein, Using clinicians’ search query data to monitor influenza epidemics, *Clinical Infectious Diseases* 59 (10) (2014) 1446–1450.
- [702] G. Eysenbach, Infodemiology: Tracking flu-related searches on the web for syndromic surveillance, *AMIA Annual Symposium Proceedings* 2006 (2006) 244–248.
- [703] P. M. Polgreen, Y. Chen, D. M. Pennock, F. D. Nelson, R. A. Weinstein, Using Internet searches for influenza surveillance, *Clinical Infectious Diseases* 47 (11) (2008) 1443–1448.
- [704] C. Pelat, C. Turbelin, A. Bar-Hen, A. Flahault, A.-J. Valleron, More diseases tracked by using Google Trends, *Emerging Infectious Diseases* 15 (8) (2009) 1327–1328.
- [705] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, L. Brilliant, Detecting influenza epidemics using search engine query data, *Nature* 457 (7232) (2009) 1012–1014.
- [706] S. Cook, C. Conrad, A. L. Fowlkes, M. H. Mohebbi, Assessing Google Flu Trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic, *PLoS ONE* 6 (8) (2011) e23610.
- [707] S. Goel, J. M. Hofman, S. Lahaie, D. M. Pennock, D. J. Watts, Predicting consumer behavior with Web search, *Proceedings of the National Academy of Sciences of the United States of America* 107 (41) (2010) 17486–17490.
- [708] D. R. Olson, K. J. Konty, M. Paladini, C. Viboud, L. Simonsen, Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: A comparative epidemiological study at three geographic scales, *PLoS Computational Biology* 9 (10) (2013) e1003256.
- [709] D. Lazer, R. Kennedy, G. King, A. Vespignani, The parable of Google Flu: Traps in big data analysis, *Science* 343 (6176) (2014) 1203–1205.
- [710] A. F. Dugas, M. Jalalpour, Y. Gel, S. Levin, F. Torcaso, T. Igusa, R. E. Rothman, Influenza forecasting with Google Flu Trends, *PLoS ONE* 8 (2) (2013) e56176.
- [711] M. A. Benjamin, R. A. Rigby, D. M. Stasinopoulos, Generalized autoregressive moving average models, *Journal of the American Statistical Association* 98 (461) (2003) 214–223.
- [712] O. M. Araz, D. Bentley, R. L. Muelleman, Using Google Flu Trends data in forecasting influenza-like-illness related ED visits in Omaha, Nebraska, *American Journal of Emergency Medicine* 32 (9) (2014) 1016–1023.
- [713] T. Preis, H. S. Moat, Adaptive nowcasting of influenza outbreaks using Google searches, *Royal Society Open Science* 1 (2) (2014) 140095.
- [714] S. G. Makridakis, S. C. Wheelwright, V. E. McGee, *Forecasting: Methods and Applications*, 3rd Edition, John Wiley & Sons, New York, NY, USA, 1998.

- [715] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, G. M. Ljung, *Time Series Analysis: Forecasting and Control*, 5th Edition, John Wiley & Sons, Hoboken, NJ, USA, 2015.
- [716] Q. Xu, Y. R. Gel, L. L. R. Ramirez, K. Nezafati, Q. Zhang, K.-L. Tsui, Forecasting influenza in Hong Kong with Google search queries and statistical model fusion, *PLoS ONE* 12 (5) (2017) e0176690.
- [717] Y. Teng, D. Bi, G. Xie, Y. Jin, Y. Huang, B. Lin, X. An, D. Feng, Y. Tong, Dynamic forecasting of Zika epidemics using Google Trends, *PLoS ONE* 12 (1) (2017) e0165085.
- [718] S. Yang, M. Santillana, S. C. Kou, Accurate estimation of influenza epidemics using Google search data via ARGO, *Proceedings of the National Academy of Sciences of the United States of America* 112 (47) (2015) 14473–14478.
- [719] S. Yang, S. C. Kou, F. Lu, J. S. Brownstein, N. Brooke, M. Santillana, Advances in using Internet searches to track dengue, *PLoS Computational Biology* 13 (7) (2017) e1005607.
- [720] Q. Yuan, E. O. Nsoesie, B. Lv, G. Peng, R. Chunara, J. S. Brownstein, Monitoring influenza epidemics in China with search query from Baidu, *PLoS ONE* 8 (5) (2013) e64323.
- [721] Z. Li, T. Liu, G. Zhu, H. Lin, Y. Zhang, J. He, A. Deng, Z. Peng, J. Xiao, S. Rutherford, R. Xie, W. Zeng, X. Li, W. Ma, Dengue Baidu Search Index data can improve the prediction of local dengue epidemic: A case study in Guangzhou, China, *PLoS Neglected Tropical Diseases* 11 (3) (2017) e0005354.
- [722] C. Chew, G. Eysenbach, Pandemics in the age of Twitter: Content analysis of tweets during the 2009 H1N1 outbreak, *PLoS ONE* 5 (11) (2010) e14118.
- [723] A. Culotta, Towards detecting influenza epidemics by analyzing Twitter messages, in: *Proceedings of the First Workshop on Social Media Analytics, SOMA'10*, ACM Press, New York, NY, USA, 2010, pp. 115–122.
- [724] E. Aramaki, S. Maskawa, M. Morita, Twitter catches the flu: Detecting influenza epidemics using Twitter, in: *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP'11*, Association for Computational Linguistics, Stroudsburg, PA, USA, 2011, pp. 1568–1576.
- [725] N. Cristianini, J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge University Press, New York, NY, USA, 2000.
- [726] A. Signorini, A. M. Segre, P. M. Polgreen, The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic, *PLoS ONE* 6 (5) (2011) e19467.
- [727] M. Salathé, S. Khandelwal, Assessing vaccination sentiments with online social media: Implications for infectious disease dynamics and control, *PLoS Computational Biology* 7 (10) (2011) e1002199.
- [728] A. Lamb, M. J. Paul, M. Dredze, Separating fact from fear: Tracking flu infections on Twitter, in: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Stroudsburg, PA, USA*, 2013, pp. 789–795.
- [729] D. A. Broniatowski, M. J. Paul, M. Dredze, National and local influenza surveillance through Twitter: An analysis of the 2012-2013 influenza epidemic, *PLoS ONE* 8 (12) (2013) e83672.
- [730] E.-K. Kim, J. H. Seok, J. S. Oh, H. W. Lee, K. H. Kim, Use of hangeul Twitter to track and predict human influenza infection, *PLoS ONE* 8 (7) (2013) e69305.
- [731] R. Tibshirani, Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society. Series B (Methodological)* 58 (1) (1996) 267–288.
- [732] M. J. Paul, M. Dredze, D. Broniatowski, Twitter improves influenza forecasting, *PLoS Currents* 6 (2014) 39911.
- [733] A. A. Aslam, M.-H. Tsou, B. H. Spitzberg, L. An, J. M. Gawron, D. K. Gupta, K. M. Peddeccord, A. C. Nagel, C. Allen, J.-A. Yang, S. Lindsay, The reliability of tweets as a supplementary method of seasonal influenza surveillance, *Journal of Medical Internet Research* 16 (11) (2014) e250.
- [734] A. Culotta, Estimating county health statistics with Twitter, in: *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems, CHI'14*, ACM Press, New York, NY, USA, 2014, pp. 1335–1344.
- [735] L. Chen, K. T. Hossain, P. Butler, N. Ramakrishnan, B. A. Prakash, Syndromic surveillance of Flu on Twitter using weakly supervised temporal topic models, *Data Mining and Knowledge Discovery* 30 (3) (2016) 681–710.
- [736] I. Kagashe, Z. Yan, I. Suheriani, Enhancing seasonal influenza surveillance: Topic analysis of widely used medicinal drugs using Twitter data, *Journal of Medical Internet Research* 19 (9) (2017) e315.
- [737] K. Lee, A. Agrawal, A. Choudhary, Real-time disease surveillance using Twitter data: Demonstration on flu and cancer, in: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'13*, ACM Press, New York, NY, USA, 2013, pp. 1474–1477.
- [738] M. Dredze, R. Cheng, M. J. Paul, D. Broniatowski, HealthTweets.org: A platform for public health surveillance using Twitter, in: *Proceedings of the First International Workshop on the World Wide Web and Public Health Intelligence, W3PHI 2014*, AAAI Press, Palo Alto, CA, USA, 2014, pp. 593–596.
- [739] F. J. Grajales III, S. Sheps, K. Ho, H. Novak-Lauscher, G. Eysenbach, Social media: A review and tutorial of applications in medicine and health care, *Journal of Medical Internet Research* 16 (2) (2014) e13.
- [740] M. Santillana, A. T. Nguyen, M. Dredze, M. J. Paul, E. O. Nsoesie, J. S. Brownstein, Combining search, social media, and traditional data sources to improve influenza surveillance, *PLoS Computational Biology* 11 (10) (2015) e1004513.
- [741] D. J. McIver, J. S. Brownstein, Wikipedia usage estimates prevalence of influenza-like illness in the United States in near real-time, *PLoS Computational Biology* 10 (4) (2014) e1003581.
- [742] N. Generous, G. Fairchild, A. Deshpande, S. Y. Del Valle, R. Priedhorsky, Global disease monitoring and forecasting with Wikipedia, *PLoS Computational Biology* 10 (11) (2014) e1003892.
- [743] K. S. Hickmann, G. Fairchild, R. Priedhorsky, N. Generous, J. M. Hyman, A. Deshpande, S. Y. Del Valle, Forecasting the 2013-2014 influenza season using Wikipedia, *PLoS Computational Biology* 11 (5) (2015) e1004239.
- [744] Y. Zha, T. Zhou, C. Zhou, Unfolding large-scale online collaborative human dynamics, *Proceedings of the National Academy of Sciences of the United States of America* 113 (51) (2016) 14627–14632.

- [745] G. Fairchild, S. Y. Del Valle, L. De Silva, A. M. Segre, Eliciting disease data from Wikipedia articles, in: Proceedings of the Ninth International AAAI Conference on Web and Social Media, ICWSM'15, AAAI Press, Palo Alto, CA, USA, 2015, pp. 26–33.
- [746] R. Friedhorsky, D. Osthuis, A. R. Daughton, K. R. Moran, N. Generous, G. Fairchild, A. Deshpande, S. Y. Del Valle, Measuring global disease with Wikipedia: Success, failure, and a research agenda, in: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW'17, ACM Press, New York, NY, USA, 2017, pp. 1812–1834.
- [747] J. D. Sharpe, R. S. Hopkins, R. L. Cook, C. W. Striley, Evaluating Google, Twitter, and Wikipedia as tools for influenza surveillance using Bayesian change point analysis: A comparative analysis, *JMIR Public Health and Surveillance* 2 (2) (2016) e161.
- [748] D. Barry, J. A. Hartigan, A Bayesian analysis for change point problems, *Journal of the American Statistical Association* 88 (421) (1993) 309–319.
- [749] A. J. Tatem, Y. Qiu, D. L. Smith, O. Sabot, A. S. Ali, B. Moonen, The use of mobile phone data for the estimation of the travel patterns and imported Plasmodium falciparum rates among Zanzibar residents, *Malaria Journal* 8 (2009) 287.
- [750] A. Wesolowski, N. Eagle, A. J. Tatem, D. L. Smith, A. M. Noor, R. W. Snow, C. O. Buckee, Quantifying the impact of human mobility on malaria, *Science* 338 (6104) (2012) 267–270.
- [751] A. Wesolowski, W. P. O'Meara, A. J. Tatem, S. Ndege, N. Eagle, C. O. Buckee, Quantifying the impact of accessibility on preventive healthcare in sub-Saharan Africa using mobile phone data, *Epidemiology* 26 (2) (2015) 223–228.
- [752] A. J. Tatem, Z. Huang, C. Narib, U. Kumar, D. Kandula, D. K. Pindolia, D. L. Smith, J. M. Cohen, B. Graupe, P. Uusiku, C. Lourenço, Integrating rapid risk mapping and mobile phone call record data for strategic malaria elimination planning, *Malaria Journal* 13 (2014) 52.
- [753] A. Wesolowski, G. Stresman, N. Eagle, J. Stevenson, C. Owaga, E. Marube, T. Bousema, C. Drakeley, J. Cox, C. O. Buckee, Quantifying travel behavior for infectious disease research: A comparison of data from surveys and mobile phones, *Scientific Reports* 4 (2014) 5678.
- [754] X.-Y. Yan, X.-P. Han, B.-H. Wang, T. Zhou, Diversity of individual mobility patterns and emergence of aggregated scaling laws, *Scientific Reports* 3 (2013) 2678.
- [755] M. Tizzoni, P. Bajardi, A. Decuyper, G. K. K. King, C. M. Schneider, V. Blondel, Z. Smoreda, M. C. González, V. Colizza, On the use of human mobility proxies for modeling epidemics, *PLoS Computational Biology* 10 (7) (2014) e1003716.
- [756] A. Wesolowski, C. Metcalf, N. Eagle, J. Kombich, B. T. Grenfell, O. N. Bjørnstad, J. Lessler, A. J. Tatem, C. O. Buckee, Quantifying seasonal population fluxes driving rubella transmission dynamics using mobile phone data, *Proceedings of the National Academy of Sciences of the United States of America* 112 (35) (2015) 11114–11119.
- [757] A. Wesolowski, T. Qureshi, M. F. Boni, P. R. Sundsøy, M. A. Johansson, S. B. Rasheed, K. Engø-Monsen, C. O. Buckee, Impact of human mobility on the emergence of dengue epidemics in Pakistan, *Proceedings of the National Academy of Sciences of the United States of America* 112 (38) (2015) 11887–11892.
- [758] J. Lourenço, M. Recker, The 2012 madeira dengue outbreak: Epidemiological determinants and future epidemic potential, *PLoS Neglected Tropical Diseases* 8 (8) (2014) e3083.
- [759] L. Bengtsson, J. Gaudart, X. Lu, S. Moore, E. Wetter, K. Sallah, S. Rebaudet, R. Piarroux, Using mobile phone data to predict the spatial spread of cholera, *Scientific Reports* 5 (2015) 8923.
- [760] F. Finger, T. Genolet, L. Mari, G. C. de Magny, N. M. Manga, A. Rinaldo, E. Bertuzzo, Mobile phone data highlights the role of mass gatherings in the spreading of cholera outbreaks, *Proceedings of the National Academy of Sciences of the United States of America* 113 (23) (2016) 6421–6426.
- [761] Y.-A. de Montjoye, Z. Smoreda, R. Trinquart, C. Ziemlicki, V. D. Blondel, D4D-Senegal: The second mobile phone data for development challenge, *arXiv:1407.4885*, 2014.
- [762] A. M. Tompkins, N. McCreesh, Migration statistics relevant for malaria transmission in Senegal derived from mobile phone data and used in an agent-based migration model, *Geospatial Health* 11 (1 Suppl) (2016) 408.
- [763] J. de Monasterio, A. Salles, C. Lang, D. Weinberg, M. Minnoni, M. Travizano, C. Sarraute, Analyzing the spread of chagas disease with mobile phone data, in: 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM'16, IEEE Press, 2016, pp. 607–612.
- [764] A. Wesolowski, E. zu Erbach-Schoenberg, A. J. Tatem, C. Lourenço, C. Viboud, V. Charu, N. Eagle, K. Engø-Monsen, T. Qureshi, C. O. Buckee, C. J. E. Metcalf, Multinational patterns of seasonal asymmetry in human movement influence infectious disease dynamics, *Nature Communications* 8 (2017) 2069.
- [765] C. Panigutti, M. Tizzoni, P. Bajardi, Z. Smoreda, V. Colizza, Assessing the use of mobile phone data to describe recurrent mobility patterns in spatial epidemic models, *Royal Society Open Science* 4 (5) (2017) 160950.
- [766] H. W. Hethcote, The mathematics of infectious diseases, *SIAM Review* 42 (4) (2000) 599–653.
- [767] L. Mari, M. Gatto, M. Ciddio, E. D. Dia, S. H. Sokolow, G. A. de Leo, R. Casagrandi, Big-data-driven modeling unveils country-wide drivers of endemic schistosomiasis, *Scientific Reports* 7 (2017) 489.
- [768] A. Wesolowski, C. O. Buckee, K. Engø-Monsen, C. J. E. Metcalf, Connecting mobility to infectious diseases: The promise and limits of mobile phone data, *Journal of Infectious Diseases* 214 (suppl\_4) (2016) S414–S420.
- [769] K. H. Jones, H. Daniels, S. Heys, D. V. Ford, Challenges and potential opportunities of mobile phone call detail records in health research: Review, *JMIR Mhealth Uhealth* 6 (7) (2018) e161.
- [770] Q. Huang, Y. Xiao, Geographic situational awareness: Mining tweets for disaster preparedness, emergency response, impact, and recovery, *ISPRS International Journal of Geo-Information* 4 (3) (2015) 1549–1568.
- [771] M. Yu, C. Yang, Y. Li, Big data in natural disaster management: A review, *Geosciences* 8 (5) (2018) 165.
- [772] M. Imran, C. Castillo, F. Diaz, S. Vieweg, Processing social media messages in mass emergency: A survey, *ACM Computing Surveys* 47 (4) (2015) 67.
- [773] N. Bellomo, D. Clarke, L. Gibelli, P. Townsend, B. Vreugdenhil, Human behaviours in evacuation crowd dynamics: From modelling to “big data” toward crisis management, *Physics of Life Reviews* 18 (2016) 1–21.
- [774] X. Zhou, Y. Shi, X. Deng, Y. Deng, D-DEMATEL: A new method to identify critical success factors in emergency management, *Safety Science* 91 (2017) 93–104.
- [775] Y. Han, D. Yong, A hybrid intelligent model for assessment of critical success factors in high-risk emergency system, *Journal of Ambient*

- Intelligence and Humanized Computing 9 (6) (2018) 1933–1953.
- [776] S. Plank, Rapid damage assessment by means of multi-temporal SAR—A comprehensive review and outlook to Sentinel-1, *Remote Sensing* 6 (6) (2014) 4870–4906.
- [777] F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, C. Ratti, Real-time urban monitoring using cell phones: A case study in Rome, *IEEE Transactions on Intelligent Transportation Systems* 12 (1) (2011) 141–151.
- [778] C. Castillo, *Big Crisis Data: Social Media in Disasters and Time-Critical Situations*, Cambridge University Press, New York, NY, USA, 2016.
- [779] S. Voigt, T. Kemper, T. Riedlinger, R. Kiefl, K. Scholte, H. Mehl, Satellite image analysis for disaster and crisis-management support, *IEEE Transactions on Geoscience and Remote Sensing* 45 (6) (2007) 1520–1528.
- [780] K. E. Joyce, S. E. Belliss, S. V. Samsonov, S. J. Mcneill, P. J. Glassey, A review of the status of satellite remote sensing and image processing techniques for mapping natural hazards and disasters, *Progress in Physical Geography: Earth and Environment* 33 (2) (2009) 183–207.
- [781] D. Massonnet, M. Rossi, C. Carmona, F. Adragna, G. Peltzer, K. Feigl, T. Rabaut, The displacement field of the landers earthquake mapped by radar interferometry, *Nature* 364 (6433) (1993) 138–142.
- [782] H. Miura, S. Midorikawa, Updating GIS building inventory data using high-resolution satellite images for earthquake damage assessment: Application to metro Manila, Philippines, *Earthquake Spectra* 22 (1) (2006) 151–168.
- [783] C. Marin, F. Bovolo, L. Bruzzone, Building change detection in multitemporal very high resolution SAR images, *IEEE Transactions on Geoscience and Remote Sensing* 53 (5) (2015) 2664–2682.
- [784] C. Corbane, D. Carrion, G. Lemoine, M. Broglia, Comparison of damage assessment maps derived from very high spatial resolution satellite and aerial imagery produced for the Haiti 2010 Earthquake, *Earthquake Spectra* 27 (S1) (2011) S199–S218.
- [785] P. Uprety, F. Yamazaki, Use of high-resolution SAR intensity images for damage detection from the 2010 Haiti Earthquake, in: 2012 IEEE International Geoscience and Remote Sensing Symposium, IEEE Press, 2012, pp. 6829–6832.
- [786] D. Ehrlich, X. Blaes, P. Soille, Extracting building stock information from optical satellite imagery for mapping earthquake exposure and its vulnerability, *Natural Hazards* 68 (1) (2013) 79–95.
- [787] J. Tian, A. A. Nielsen, P. Reinartz, Building damage assessment after the earthquake in Haiti using two post-event satellite stereo imagery and DSMs, *International Journal of Image and Data Fusion* 6 (2) (2015) 155–169.
- [788] Y.-A. Liou, S. K. Kar, L. Chang, Use of high-resolution FORMOSAT-2 satellite images for post-earthquake disaster assessment: A study following the 12 May 2008 Wenchuan Earthquake, *International Journal of Remote Sensing* 31 (13) (2010) 3355–3368.
- [789] X. Tong, Z. Hong, S. Liu, X. Zhang, H. Xie, Z. Li, S. Yang, W. Wang, F. Bao, Building-damage detection using pre- and post-seismic high-resolution satellite stereo imagery: A case study of the May 2008 Wenchuan Earthquake, *ISPRS Journal of Photogrammetry and Remote Sensing* 68 (2012) 13–27.
- [790] R. G. Congalton, A review of assessing the accuracy of classifications of remotely sensed data, *Remote Sensing of Environment* 37 (1) (1991) 35–46.
- [791] D. Jin, X. Wang, A. Dou, Y. Dong, Post-earthquake building damage assessment in Yushu using airborne SAR imagery, *Earthquake Science* 24 (5) (2011) 463–473.
- [792] L. Shi, W. Sun, J. Yang, P. Li, L. Lu, Building collapse assessment by the use of postearthquake Chinese VHR airborne SAR, *IEEE Geoscience and Remote Sensing Letters* 12 (10) (2015) 2021–2025.
- [793] J. Schmidhuber, Deep learning in neural networks: An overview, *Neural Networks* 61 (2015) 85–117.
- [794] A. J. Cooner, Y. Shao, J. B. Campbell, Detection of urban damage using remote sensing and machine learning algorithms: Revisiting the 2010 Haiti Earthquake, *Remote Sensing* 8 (10) (2016) 868.
- [795] G. Sun, Y. Hao, J. Rong, S. Shi, J. Ren, Combined deep learning and multiscale segmentation for rapid high resolution damage mapping, in: 2017 IEEE International Conference on Internet of Things, IEEE Press, 2017, pp. 1101–1105.
- [796] A. Fujita, K. Sakurada, T. Imaizumi, R. Ito, S. Hikosaka, R. Nakamura, Damage detection from aerial images via convolutional neural networks, in: 2017 Fifteenth IAPR International Conference on Machine Vision Applications, IEEE Press, 2017, pp. 5–8.
- [797] Y. Bai, E. Mas, S. Koshimura, Towards operational satellite-based damage-mapping using U-Net Convolutional Network: A case study of 2011 Tohoku Earthquake-Tsunami, *Remote Sensing* 10 (10) (2018) 1626.
- [798] J. Zhao, F. Ding, Z. Wang, J. Ren, J. Zhao, Y. Wang, X. Tang, Y. Wang, J. Yao, Q. Li, A rapid public health needs assessment framework for after major earthquakes using high-resolution satellite imagery, *International Journal of Environmental Research and Public Health* 15 (6) (2018) 1111.
- [799] H. R. Ranjbar, A. A. Ardalan, H. Dehghani, M. R. Saradjian, Using high-resolution satellite imagery to provide a relief priority map after earthquake, *Natural Hazards* 90 (3) (2018) 1087–1113.
- [800] J. A. Quinn, M. M. Nyhan, C. Navarro, D. Coluccia, L. Bromley, M. Luengo-Oroz, Humanitarian applications of machine learning with remote-sensing data: Review and case study in refugee settlement mapping, *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 376 (2128) (2018) 20170363.
- [801] V. Klemas, Remote sensing of floods and flood-prone areas: An overview, *Journal of Coastal Research* 31 (4) (2015) 1005–1013.
- [802] J. Sanyal, X. X. Lu, Application of remote sensing in flood management with special reference to monsoon Asia: A review, *Natural Hazards* 33 (2) (2004) 283–301.
- [803] P. A. Brivio, R. Colombo, M. Maggi, R. Tomasoni, Integration of remote sensing data and GIS for accurate mapping of flooded areas, *International Journal of Remote Sensing* 23 (3) (2002) 429–441.
- [804] T. D. Groeve, Flood monitoring and mapping using passive microwave remote sensing in Namibia, *Geomatics, Natural Hazards and Risk* 1 (1) (2010) 19–35.
- [805] S. Skakun, N. Kussul, A. Shelestov, O. Kussul, Flood hazard and flood risk assessment using a time series of satellite images: A case study in Namibia, *Risk Analysis* 34 (8) (2014) 1521–1537.
- [806] L. Giustarini, M. Chini, R. Hostache, F. Pappenberger, P. Matgen, Flood hazard mapping combining hydrodynamic modeling and multi annual remote sensing data, *Remote Sensing* 7 (10) (2015) 14200–14226.
- [807] Y.-J. Kwak, Nationwide flood monitoring for disaster risk reduction using multiple satellite data, *ISPRS International Journal of Geo-*

Information 6 (7) (2017) 203.

- [808] M. S. Rahman, L. Di, The state of the art of spaceborne remote sensing in flood management, *Natural Hazards* 85 (2) (2016) 1223–1248.
- [809] A. D'Addabbo, A. Refice, D. Capolongo, G. Pasquariello, S. Manfreda, Data fusion through bayesian methods for flood monitoring from remotely sensed data, in: A. Refice, A. D'Addabbo, D. Capolongo (Eds.), *Flood Monitoring through Remote Sensing*, Springer, Cham, Switzerland, 2018, pp. 181–208.
- [810] I. T. Ekeu-wei, G. A. Blackburn, Applications of open-access remotely sensed data for flood modelling and mapping in developing regions, *Hydrology* 5 (3) (2018) 39.
- [811] A. Refice, A. D'Addabbo, D. Capolongo, *Flood Monitoring through Remote Sensing*, Springer, Cham, Switzerland, 2018.
- [812] P. W. Gething, A. J. Tatem, Can mobile phone data improve emergency response to natural disasters?, *PLoS Medicine* 8 (8) (2011) e1001085.
- [813] Z.-Q. Jiang, W.-J. Xie, M.-X. Li, B. Podobnik, W.-X. Zhou, H. E. Stanley, Calling patterns in human communication dynamics, *Proceedings of the National Academy of Sciences of the United States of America* 110 (5) (2013) 1600–1605.
- [814] J. P. Bagrow, D. Wang, A.-L. Barabási, Collective response of human populations to large-scale emergencies, *PLoS ONE* 6 (3) (2011) e17680.
- [815] B. Moumni, V. Frias-Martinez, E. Frias-Martinez, Characterizing social response to urban earthquakes using cell-phone network data: The 2012 Oaxaca earthquake, in: *Proceedings of the 2013 ACM Conference on Pervasive and Ubiquitous Computing Adjunct Publication, UbiComp'13 Adjunct*, ACM Press, New York, NY, USA, 2013, pp. 1199–1208.
- [816] L. Gao, C. Song, Z. Gao, A.-L. Barabási, J. P. Bagrow, D. Wang, Quantifying information flow during emergencies, *Scientific Reports* 4 (2014) 3997.
- [817] X. Yu, T. Pei, K. Gai, L. Guo, Analysis on urban collective call behavior to earthquake, in: *2015 IEEE 17th International Conference on High Performance Computing and Communications*, IEEE Press, 2015, pp. 1302–1307.
- [818] D. Pastor-Escuredo, A. Morales-Guzmán, Y. Torres-Fernández, J.-M. Bauer, A. Wadhwa, C. Castro-Correa, L. Romanoff, J. G. Lee, A. Rutherford, V. Frias-Martinez, N. Oliver, E. Frias-Martinez, M. Luengo-Oroz, Flooding through the lens of mobile phone activity, in: *IEEE Global Humanitarian Technology Conference, GHTC 2014*, IEEE Press, 2014, pp. 279–286.
- [819] L. Hong, M. Lee, A. Mashhadni, V. Frias-Martinez, Towards understanding communication behavior changes during floods using cell phone data, in: *Proceedings of the 10th International Conference on Social Informatics*, Springer, Cham, Switzerland, 2018, pp. 97–107.
- [820] A. Dobra, N. E. Williams, N. Eagle, Spatiotemporal detection of unusual human population behavior using mobile phone data, *PLoS ONE* 10 (3) (2015) e0120449.
- [821] D. Gundogdu, O. D. Incel, A. A. Salah, B. Lepri, Countrywide arrhythmia: Emergency event detection using mobile phone data, *EPJ Data Science* 5 (2016) 25.
- [822] L. Bengtsson, X. Lu, A. Thorson, R. Garfield, J. von Schreeb, Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: A post-earthquake geospatial study in Haiti, *PLoS Medicine* 8 (8) (2011) e1001083.
- [823] X. Lu, L. Bengtsson, P. Holme, Predictability of population displacement after the 2010 Haiti earthquake, *Proceedings of the National Academy of Sciences of the United States of America* 109 (29) (2012) 11576–11581.
- [824] D. Y. Kenett, J. Portugali, Population movement under extreme events, *Proceedings of the National Academy of Sciences of the United States of America* 109 (29) (2012) 11472–11473.
- [825] X. Song, Q. Zhang, Y. Sekimoto, T. Horanont, S. Ueyama, R. Shibasaki, Modeling and probabilistic reasoning of population evacuation during large-scale disaster, in: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'13*, ACM Press, New York, NY, USA, 2013, pp. 1231–1239.
- [826] X. Song, Q. Zhang, Y. Sekimoto, R. Shibasaki, Prediction of human emergency behavior and their mobility following large-scale disaster, in: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'14*, ACM Press, New York, NY, USA, 2014, pp. 5–14.
- [827] N. Bharti, X. Lu, L. Bengtsson, E. Wetter, A. J. Tatem, Remotely measuring populations during a crisis by overlaying two data sources, *International Health* 7 (2) (2015) 90–98.
- [828] R. Wilson, E. Zu Schoenberg-Elisabeth, M. Albert, D. Power, S. Tudge, M. Gonzalez, S. Guthrie, H. Chamberlain, C. Brooks, C. Hughes, L. Pitonakova, C. Buckee, X. Lu, E. Wetter, A. Tatem, L. Bengtsson, Rapid and near real-time assessments of population displacement using mobile phone data following disasters: The 2015 Nepal Earthquake, *PLoS Currents* 8 (2016) 27109.
- [829] J. Ghurye, G. Krings, V. Frias-Martinez, A framework to model human behavior at large scale during natural disasters, in: *2016 17th IEEE International Conference on Mobile Data Management, MDM'16*, IEEE Press, 2016, pp. 18–27.
- [830] X. He, Y.-R. Lin, Measuring and monitoring collective attention during shocking events, *EPJ Data Science* 6 (2017) 30.
- [831] Y. Sano, K. Yamada, H. Watanabe, H. Takayasu, M. Takayasu, Empirical analysis of collective human behavior for extraordinary events in the blogosphere, *Physical Review E* 87 (1) (2013) 012805.
- [832] Z. Wang, X. Ye, Social media analytics for natural disaster management, *International Journal of Geographical Information Science* 32 (1) (2018) 49–72.
- [833] K. Kireyev, L. Palen, K. Anderson, Applications of topics models to analysis of disaster-related Twitter data, in: *Proceedings of NIPS Workshop on Applications for Topic Models: Text and Beyond*, Vol. 1, NIPS, Inc., La Jolla, CA, USA, 2009, pp. 1–4.
- [834] M. Imran, S. Elbassuoni, C. Castillo, F. Diaz, P. Meier, Extracting information nuggets from disaster-related messages in social media, in: *Proceedings of the 10th International Conference on Information Systems for Crisis Response and Management, ISCRAM'13*, 2013, pp. 791–800.
- [835] A. Olteanu, C. Castillo, F. Diaz, S. Vieweg, Crisislex: A lexicon for collecting and filtering microblogged communications in crises, in: *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media, ICWSM'14*, AAAI Press, Palo Alto, CA, USA, 2014, pp. 376–385.
- [836] Z. Ashktorab, C. Brown, M. Nandi, A. Culotta, Tweedr: Mining Twitter to inform disaster response, in: *Proceedings of the 11th International Conference on Information Systems for Crisis Response and Management, ISCRAM'14*, 2014, pp. 354–358.
- [837] R. M. Allen, Transforming earthquake detection?, *Science* 335 (6066) (2012) 297–298.
- [838] A. Acar, Y. Muraki, Twitter for crisis communication: Lessons learned from Japan's tsunami disaster, *International Journal of Web Based*

- Communities 7 (3) (2011) 392–402.
- [839] F. Toriumi, T. Sakaki, K. Shinoda, K. Kazama, S. Kurihara, I. Noda, Information sharing on Twitter during the 2011 catastrophic earthquake, in: Proceedings of the 22nd International Conference on World Wide Web, WWW’13 Companion, ACM Press, New York, NY, USA, 2013, pp. 1025–1028.
- [840] A. Chatfield, U. Brajawidagda, Twitter tsunami early warning network: A social network analysis of Twitter information flows, in: Proceedings of the 23rd Australasian Conference on Information Systems, ACIS’12, Deakin University, Geelong, Vic., Australia, 2012, pp. 1–10.
- [841] R. Dong, L. Li, Q. Zhang, G. Cai, Information diffusion on social media during natural disasters, IEEE Transactions on Computational Social Systems 5 (1) (2018) 265–276.
- [842] T. Sakaki, M. Okazaki, Y. Matsuo, Earthquake shakes Twitter users: Real-time event detection by social sensors, in: Proceedings of the 19th International Conference on World Wide Web, WWW’10, ACM Press, New York, NY, USA, 2010, pp. 851–860.
- [843] J. Hightower, G. Borriello, Particle filters for location estimation in ubiquitous computing: A case study, in: Proceedings of the 6th International Conference on Ubiquitous Computing, UbiComp’04, Springer, Berlin, Heidelberg, 2004, pp. 88–106.
- [844] T. Sakaki, M. Okazaki, Y. Matsuo, Tweet analysis for real-time event detection and earthquake reporting system development, IEEE Transactions on Knowledge and Data Engineering 25 (4) (2013) 919–931.
- [845] P. S. Earle, D. C. Bowden, M. Guy, Twitter earthquake detection: Earthquake monitoring in a social world, Annals of Geophysics 54 (6) (2012) 708–715.
- [846] B. Robinson, R. Power, M. Cameron, A sensitive Twitter earthquake detector, in: Proceedings of the 19th International Conference on World Wide Web, WWW’13, ACM Press, New York, NY, USA, 2013, pp. 999–1002.
- [847] J.-Y. Jung, M. Moro, Multi-level functionality of social media in the aftermath of the Great East Japan Earthquake, Disasters 38 (s2) (2014) s123–s143.
- [848] S. Vieweg, A. L. Hughes, K. Starbird, L. Palen, Microblogging during two natural hazards events: What Twitter may contribute to situational awareness, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI’10, ACM Press, New York, NY, USA, 2010, pp. 1079–1088.
- [849] F. Cheong, C. Cheong, Social media data mining: A social network analysis of tweets during the Australian 2010–2011 floods, in: Proceedings of the 15th Pacific Asia Conference on Information Systems, PACIS 2011, Queensland University of Technology, Brisbane, Australia, 2011, p. 46.
- [850] J. P. de Albuquerque, B. Herfort, A. Brenning, A. Zipf, A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management, International Journal of Geographical Information Science 29 (4) (2015) 667–689.
- [851] D. Eilander, P. Trambauer, J. Wagemaker, A. van Loenen, Harvesting social media for generation of near real-time flood maps, Procedia Engineering 154 (2016) 176–183.
- [852] R. Arthur, C. A. Boulton, H. Shotton, H. T. Williams, Social sensing of floods in the UK, PLoS ONE 13 (1) (2018) e0189327.
- [853] Z. Li, C. Wang, C. T. Emrich, D. Guo, A novel approach to leveraging social media for rapid flood mapping: A case study of the 2015 South Carolina floods, Cartography and Geographic Information Science 45 (2) (2018) 97–110.
- [854] N. Tkachenko, S. Jarvis, R. Procter, Predicting floods with Flickr tags, PLoS ONE 12 (2) (2017) e0172870.
- [855] J. F. Rosser, D. Leibovici, M. Jackson, Rapid flood inundation mapping using social media, remote sensing and topographic data, Natural Hazards 87 (1) (2017) 103–120.
- [856] R.-Q. Wang, H. Mao, Y. Wang, C. Rae, W. Shaw, Hyper-resolution monitoring of urban flooding with social media and crowdsourcing data, Computers and Geosciences 111 (2018) 139–147.
- [857] C. D. Manning, H. Schütze, Foundations of Statistical Natural Language Processing, MIT Press, 1999.
- [858] T. H. Assumpção, I. Popescu, A. Jonoski, D. P. Solomatine, Citizen observations contributing to flood modelling: Opportunities and challenges, Hydrology and Earth System Sciences 22 (2) (2018) 1473–1489.
- [859] T. Preis, H. S. Moat, S. R. Bishop, P. Treleaven, H. E. Stanley, Quantifying the digital traces of Hurricane Sandy on Flickr, Scientific Reports 3 (2013) 3141.
- [860] M. G. Kendall, A new measure of rank correlation, Biometrika 30 (1–2) (1938) 81–89.
- [861] Y. Kryvasheyev, H. Chen, N. Obradovich, E. Moro, P. van Hentenryck, J. Fowler, M. Cebrrian, Rapid assessment of disaster damage using social media activity, Science Advances 2 (3) (2016) e1500779.
- [862] O. Gruebner, S. R. Lowe, M. Sykora, K. Shankardass, S. Subramanian, S. Galea, A novel surveillance approach for disaster mental health, PLoS ONE 12 (7) (2017) e0181233.
- [863] Q. Wang, J. E. Taylor, Quantifying human mobility perturbation and resilience in Hurricane Sandy, PLoS ONE 9 (11) (2014) e112608.
- [864] S. E. Middleton, L. Middleton, S. Modafferi, Real-time crisis mapping of natural disasters using social media, IEEE Intelligent Systems 29 (2) (2014) 9–17.
- [865] L. Zou, N. S. Lam, H. Cai, Y. Qiang, Mining Twitter data for improved understanding of disaster resilience, Annals of the American Association of Geographers 108 (5) (2018) 1422–1441.
- [866] Z. Tang, L. Zhang, F. Xu, H. Vo, Examining the role of social media in California’s drought risk management in 2014, Natural Hazards 79 (1) (2015) 171–193.
- [867] C. A. Boulton, H. Shotton, H. T. Williams, Using social media to detect and locate wildfires, in: Proceedings of the 19th International Conference on World Wide Web, WWW’10, ACM Press, New York, NY, USA, 2016, pp. 178–186.
- [868] L. Hong, C. Fu, P. Torrens, V. Frias-Martinez, Understanding citizens’ and local governments’ digital communications during natural disasters: The case of snowstorms, in: Proceedings of the 2017 ACM on Web Science Conference, WebSci’17, ACM Press, New York, NY, USA, 2017, pp. 141–150.
- [869] E. Ferrara, O. Varol, C. Davis, F. Menczer, A. Flammini, The rise of social bots, Communications of the ACM 59 (7) (2014) 96–104.
- [870] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, Science 359 (6380) (2018) 1146–1151.
- [871] D. Ruths, J. Pfeffer, Social media for large studies of behavior, Science 346 (6213) (2014) 1063–1064.
- [872] I. J. Good, Y. Mittal, The amalgamation and geometry of two-by-two contingency tables, Annals of Statistics 15 (2) (1987) 694–711.

- [873] M. Reinstein, J. Simanek, P. Machalek, J. Zikes, Systems, methods and computer program products for multi-resolution multi-spectral deep learning based change detection for satellite images, U.S. Patent Application, No. 15/813,455, 2019.
- [874] G. J. Duncan, K. Magnuson, Socioeconomic status and cognitive functioning: Moving from correlation to causation, Wiley Interdisciplinary Reviews Cognitive Science 3 (3) (2012) 377–386.
- [875] R. M. Bond, C. J. Fariss, J. J. Jones, A. D. I. Kramer, M. Cameron, J. E. Settle, J. H. Fowler, A 61-million-person experiment in social influence and political mobilization, Nature 489 (7415) (2012) 295–298.
- [876] K. Danaher (Ed.), 50 Years is Enough: The Case Against the World Bank and the International Monetary Fund, South End Press, Boston, MA, USA, 1994.
- [877] J. Pincus, J. A. Winters (Eds.), Reinventing the World Bank, Cornell University Press, Ithaca, NY, USA, 2002.