How could computational methods extend the border of researches in social stratification and inequality?

Written by Yizhou Ye and Yi Zhang, and theses two authors contribute equally in both the reading and writing of this literature review.

The rising of computational methods in social science is reshaping a great number of areas, including research in social stratification and mobility. In this review, we would like to summarize the innovation of computational methods in four aspects: more available data, more dimensions in measurement, the novel understanding approach of causality and mechanism, and the new goal of this area. This review is not only limited in the application of computational methods in social stratification but also shows the potential of what we could learn from other research areas.

Social stratification and inequality is one of the most important fields of both sociology and economics, which is even more important in today's world, given the condition that cyber techniques deeply interact with capitalism, increasing the inequality all around the world and all among people. However, compared with network analysis, culture study and many other areas, research in social stratification and inequality is not yet well influenced by computational social science and still highly covered by the shadow of great 20 century studies. Some people might argue that the topics and issues cared by scholars in this field are too important to be ignored and the data they need is too fundermantal to be less collected. However, in this literature review, we would elaborate how computational social science could innovate this field, including the object they could touch, the way they use it to analyze and the goal they pursue. It's inevitable to apply computational social science in this field, without which it would be almost impossible to catch the pulse of social stratification and inequality in the era of data science.

More available data and new measurement

Traditionally, researchers interested in social stratification and mobility generally use data from social surveys, when they make arguments at an individual level, or aggregate data, when they make arguments at some higher levels (DiPrete, 2020). However, more data are available today, including administrative data (Chetty et al., 2018), Internet data/ Big data (Golder & Macy, 2014), image data (Suel et al., 2019), App tracking data (Althoff et al., 2017) and sensor data (Robles-Granda, 2020), etc., which could extend people's understanding of society, including stratification and inequality.

The increase of available data makes it possible for researchers to detect minority groups and analyze the heterogeneity inside them (Lazer & Radford, 2017). It's particularly important in the field of social stratification and inequality, because of the reasons that disadvantaged groups are sometimes hard to be found in social surveys and they might be fragile in different aspects. If we would like to implement some precise policy, understanding their heterogeneity should be the precondition. The newly available data, especially big data, would also be the materials to understand what technology innovation brought to the pattern of social inequality (Edelman & Luca, 2014; Hampton, 2017), particularly compared to what surveys are doing is almost as same as decades before. Sensor data, which could be classified as one kind of behavior data, could assess both the surrounding environment and individual health status as well as lifestyle detailedly (Toomet et al., 2015). The popularization of smartphones makes them become a useful tool to record the track of people's behavior. Althoff et al. (2017) demonstrate worldwide activity inequality and its outcomes on health by collecting large scale app tracking data. Scientifically analyzing image data could also extract environmental features affecting inequality (Suel et al., 2019), rather than only qualitatively understand that (Sampson, 2012). It’s key to understand the micro process of production and maintenance of stratification and inequality.

The relationship between social network and social inequality is a classical topic in this field (Blau, 1977; DiMaggio & Garip, 2012), while the lack of available network data results in absence of rigorous empirical studies. Nowadays, we could connect network characteristics to detailed individual data with personal data footprint which could be extracted from big data (Golder & Macy, 2014). Moreover, computational methods contribute to discovering the mechanisms of how social network shapes inequality through online experiments. For instance, Shirado et al. (2019) use an online Wi-Fi sharing game to illustrate how network brokerage shapes inequality, which is very helpful for us to understand today’s community social capital’s impact on social mobility and the relationship between community social network, individual position, behaviour and the long-term outcome on resources possession .

Researchers, especially economists and sociologists, have developed a set of classical measurements of social stratification and mobility, such as income, education, occupation, household wealth, and some comprehensive indicators, like standard international socio-economic index (Ganzeboom et al. 1992). However, these traditional measurements face some shortcomings, including 1) Some information is not available in many research, like historical studies, 2) Some indicators which might perform better in subgroup studies are ignored since data only include most general variables, 3) Existing indicators, which is well developed, are limited to the traditional understanding of social inequality, leading to the ignorance of phenomena newly occurring. It’s hard to capture contemporary life if we rely on questionnaires that look like what it is in two decades ago.

Fortunately, computational methods could enrich our toolkits of studying social stratification and inequality. Historical data is key for us to understand intergenerational mobility, but how to linkage individuals from different sources (especially for the historical census) remains understudied. Abramitzky et al. (2019) develop an automatic method to link historical data. Applying such methods makes studying multigenerational mobility through historical data become possible and more accurate. Apart from data, measurement is another concern. Wage, a widely used measurement of economic status, is unlikely to be collected in much historical data. Machine learning could help us impute occupational income scores better than linear model using lasso adjustment to select appropriate interaction between several characteristics (Saavedra & Twinam, 2020). With sensor data, a study conducted by Toomet et al. (2015) gives people insights about modern life to compare time spent in mobile devices among different ethnic groups, showing the differences of ability to utilize digital resources and implying increasing inequality caused by that.

Briefly, with computational methods' assistance, we could capture far more big samples of society and we are able to understand more in history and modern life, but not only covered by the shadow of the 20th century. We could also better understand multidimensional inequality in the contemporary era.

Constructing the counterfactual and unpacking the mechanism

The counterfactual framework, which means that we should compare what happens exactly with the 'what-if' scenario to overcome the selection bias to obtain the real and net causal effect, is key for us to understand causality (Morgan & Winship, 2015). Machine learning could contribute a lot to constructing a virtual reference group (Molina & Garip, 2019), the comparison between which and reality could also open the black box, figuring out what exactly leads to inequality but under the cover of the so-called fair institution (Berk et al. 2018; Kleinberg et al., 2017). Based on the simulation, researchers could construct a world that all conditions are the same as reality but assume there is no discrimination in a specific characteristic. The difference between this constructed world and reality could be interpreted as the consequences of discrimination and researchers could further calculate the benefit if the obstacle resulting from discrimination is removed (Chetty et al., 2020). Moreover, as an extension of the counterfactual framework's application, the heterogeneous treatment effect, which is useful for researchers to understand the moderating effect of context and traditional dealt with basic matching tools (Xie et al., 2012), could also be excellently solved by machine learning (Athey & Imbens, 2019). Though studies listed above are mostly not in the area of social stratification and mobility, we could obviously learn a lot from them to enhance the causality in this field. With simulation, we could estimate the degree of consequences of discrimination; with more reliable artificial counter groups, we could rigorously examine both inequalities hidden by selection and investigate how the heterogeneous context moderates the effect of some variables.

Besides understanding causality in the counterfactual framework, which could be understood as an experimental approach, the analytical approach, which highlights the social mechanism, also benefits from computational methods (Hedström & Ylikoski, 2010). In this tradition, what researchers should focus on is individual action rather than statistical significance at the group level, implying that collection consequences are not the simple aggregation of individual action (Page, 2015). The class theory is used as an example to illustrate that if we only pay attention to macro explanation, then we could not understand the inner mechanism, what science exactly should do (Hedström & Swedberg, 1998). They appeal to apply agent-based modeling as the solution to help us investigate mechanisms (Hedström & Ylikoski, 2010). Nevertheless, as the field of computational methods is dominated by data science, ABM is less developed these years.

The investigation of social mechanisms could also benefit from the development of today's computational social science, including providing more possible information to construct ABM and large-scale experiments, etc. (Keuschnigg et al., 2018). Besides the ABM approach, computational social science also makes abundant features available now as said above, serving as an amplifier to complement probability-based explanation chains in previous mechanism analysis. The idea of the analytical approach might also contribute to the understanding of the neural network by treating it as layers of actions. It's essential to figure out chains producing inequality as detailed as possible if we would like to make social study helpful for policy makers, in both precise and efficient aims, with which people could find the way to break the reproduction of inequality.

Computational methods also contribute to the development of classic causal inference. PSM, drawing the idea of the counterfactual framework, traditionally tends to use logistic regression to impute propensity score. Recent works begin to apply machine learning methods in the process such as neural network and decision trees (Molina & Garip, 2019). The synthetic control method developed by Abadie and Gardeazabal (2003) is regarded as an important development of DID. Doudchenko and Imbens (2016) apply LASSO and elastic net in calculating weights for the synthetic control method, which can perform better in settings with a large number of units. For the instrumental variable method, Belloni et al. (2012) apply machine learning in forming first-stage predictions and estimating optimal instruments. In simulation experiments, their LASSO-IV estimators perform better than recently advocated instrument robust procedures. All these researches show us the impact brought by computational social science is not only the revolutionary innovation of research methodology, but also the pragmatic improvement of performance of existing methods and design.

Overall, thanks to the rapid development of computational social science, we could enhance the robustness of research in social stratification and mobility and create the space to understand the chain and inner mechanism of inequality (re)production, bridging micro behaviors and macro phenomena.

From identification to prediction

The traditional goal of social science researches, including social stratification and inequality, is mainly identification, which means that researchers try to figure out the real relationship between some determinants and the social consequences in social stratification (Grusky, 2014). Specifically, in the field of social stratification, researchers also try to build up a comprehensive model to explain the core pattern of social stratification and mobility, generally using the structural equation model (Blau & Duncan, 1967; Sewell et al. 1970). Researchers also try to develop some general laws of inequality (Piketty & Goldhammer, 2014). No matter comprehensive models or generalizable rules, researchers seek to answer the question about social stratification in the way of identification indeed, implying that their goal is to figure out exact determinants of social stratification and mobility and to answer what causes inequality.

The computational methods make it possible to answer this question in another way, from identification to prediction. The previous one majorly helps us to reason the origins of social inequality, while the latter one aims to solve this question, even in a way without a deep understanding of the inner mechanism. Machine learning has been widely used by various social researches, especially for prediction (Molina & Garip, 2019), which is proposed to be criteria to strengthen the scientificity of sociology and other social sciences (Watts, 2014). Matthew J. Salganik et al. (2020) make an attempt to predict children who, individually, would suffer some disadvantaged life outcomes, most of which are generally treated as classical concerns in social stratification and inequality research. Though the mass collaboration didn't lead to ideal results and the prediction is not very accurate, this study enlights people that we could take a modern mass collaboration organizational way to precisely predict individual outcomes and implement precise aid. Prediction is also essential for society. Traditional researches only explain the reason after things happen, but prediction could help prevent potential disadvantaged situations from occurring, avoiding our society falling into danger.

Besides, computational methods could also serve as the tool to generalize sample data to whole data in an individual prediction approach but not a statistical inference approach, providing us a more detailed and effective way to understand the complex picture, especially in places lacking resources to launch a qualified social survey (Blumenstock et al., 2015; Jean et al., 2016), which is generally the disadvantaged target in social stratification studies. For instance, Blumenstock et al. (2015) apply mobile phone metadata to construct the distribution of wealth of an entire nation and to infer the asset distribution of microregions, which is known as data augmentation in data science. Jean et al. (2016) go beyond previous night light data, which is widely used but inaccurate among poor areas. They combine satellite imagery and machine learning to estimate the consumption expenditure and asset wealth of some African countries. The features of those technologies are not only a supplement to social surveys in less-developed countries, the real-time feature can make it possible to estimate short-term changes in wealth and poverty, which can benefit fighting poverty (Blumenstock, 2016). Dong et al. (2019) also explore another data source that can be utilized to generalize real-time data. They apply restaurant data (from Chinese online platform *dianping*) in predicting the neighborhood's socioeconomic status. Though China does not lack survey data, large scale neighborhood surveys have never been implemented in many cities, which means their method can be applied to cities with unavailable data and benefit urban governance.

We should also pay attention to another issue, the reflexivity problem (Merton & Merton, 1968, Edelmann et al., 2020), which means our prediction and the actions based on our prediction would lead to a new situation. Hence, ideally, experiments should be implemented before the targeted policy is widely applied. Given the limitation of experiments in social science, modeling, especially agent-based modeling (Bonabeau, 2002) is an appropriate way to assess the potential outcome of precise policy.

To sum up, computational methods make it possible to precisely predict but not only figure out laws, making research in social stratification and inequality more usable in real policy making.

In conclusion, computational methods could enrich research in social stratification and inequality through the innovation of data, analytical approach, and goal. With novel new data, we could extend both the research object and dimension to understand; with computational tools, we could unpack social mechanism and make rigorous causal inference better; and with the nature of computational science, we could pursue not only explanation but prediction to both enhance the scientificity of researches in this field and increase its social impact.

All these are hard for traditional solutions to answer, implying that if we stick ourselves in existing approaches, what is limited is not only the performance of our answers but also the raising of new questions in the contemporary era.

Definitely, there are several shortcomings of computational methods. If we only rely on the data obtained from the internet, personal characteristics would be missed. If we only focus on the prediction, indeed we would be confused about what kind of actions we should take to those disadvantaged targets. Hence, it would be particularly important to connect traditional approaches and computational ones. With detailed social survey, researchers could handle abundant personal information and computational approach could link survey data with other sources, like mobile devices use, internet footprint, etc. to answer whether or how cyber technique changes individual social status. There are several ways to connect these two approaches together, including but not limited to using computational methods to analyze survey data to figure out deeper or further mechanism, using survey data to simulate social mobility, using big data characteristics to infer survey data to build up a bigger picture of social stratification, etc.

Overall, the development of computational methods promises a fruitful direction of research in social stratification and inequality. We could learn a lot from both the computational methods themselves and researches in other fields applying computational methods.

Reference

Abadie, A., & Gardeazabal, J. (2003). The economic costs of conflict: A case study of the Basque Country. American economic review, 93(1), 113-132.

Abramitzky, R., Boustan, L. P., Eriksson, K., Feigenbaum, J. J., & Pérez, S. (2019). Automated Linking of Historical Data (No. w25825). National Bureau of Economic Research. https://doi.org/10.3386/w25825

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. Annual Review of Economics, 11, 685-725.

Althoff, T., Sosič, R., Hicks, J. et al. Large-scale physical activity data reveal worldwide activity inequality. Nature 547, 336–339 (2017). https://doi.org/10.1038/nature23018

Belloni, A., Chen, D., Chernozhukov, V., & Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. Econometrica, 80(6), 2369-2429.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2018). Fairness in criminal justice risk assessments: The state of the art. Sociological Methods & Research, 0049124118782533.

Blau, P. M., & Duncan, O. D. (1967). The American occupational structure.

Grusky, D. (2014). Social Stratification. New York: Routledge, https://doi.org/10.4324/9780429494642

Blau, P. M. (1977). Inequality and heterogeneity: A primitive theory of social structure (Vol. 7). New York: Free Press.

Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. Science, 350(6264), 1073-1076.

Blumenstock, J. E. (2016). Fighting poverty with data. Science, 353(6301), 753–754. https://doi.org/10.1126/science.aah5217

Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. Proceedings of the national academy of sciences, 99(suppl 3), 7280-7287.

Chetty, R., Friedman, J. N., Hendren, N., Jones, M. R., & Porter, S. R. (2018). The opportunity atlas: Mapping the childhood roots of social mobility (No. w25147). National Bureau of Economic Research.

Chetty, R., Friedman, J. N., Saez, E., Turner, N., & Yagan, D. (2020). Income segregation and intergenerational mobility across colleges in the united states. The Quarterly Journal of Economics, 135(3), 1567-1633.

Doudchenko, N., & Imbens, G. W. (2016). Balancing, regression, difference-in-differences and synthetic control methods: A synthesis (No. w22791). National Bureau of Economic Research.

DiMaggio, P., & Garip, F. (2012). Network effects and social inequality. Annual review of sociology, 38, 93-118.

DiPrete, T. A. (2020). The Impact of Inequality on Intergenerational Mobility. Annual Review of Sociology, 46.

Dong, L., Ratti, C., & Zheng, S. (2019). Predicting neighborhoods' socioeconomic attributes using restaurant data. *Proceedings of the National Academy of Sciences*, *116*(31), 15447. https://doi.org/10.1073/pnas.1903064116

Edelman, B. G., & Luca, M. (2014). Digital discrimination: The case of Airbnb. com. Harvard Business School NOM Unit Working Paper, (14-054).

Edelmann, A., Wolff, T., Montagne, D., & Bail, C. A. (2020). Computational Social Science and Sociology. Annual Review of Sociology, 46.

Ganzeboom, H. B., De Graaf, P. M., & Treiman, D. J. (1992). A standard international socio-economic index of occupational status. Social science research, 21(1), 1-56.

Golder, S. A., & Macy, M. W. (2014). Digital footprints: Opportunities and challenges for online social research. Annual Review of Sociology, 40, 129-152.

Hampton, K. N. (2017). Studying the digital: Directions and challenges for digital methods. Annual Review of Sociology, 43, 167-188.

Hedström, P., & Swedberg, R. (1998). SOCIAL MECHANISMS.

Hedström, P., & Ylikoski, P. (2010). Causal mechanisms in the social sciences. Annual review of sociology, 36.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, *353*(6301), 790–794. https://doi.org/10.1126/science.aaf7894

Keuschnigg, M., Lovsjö, N., & Hedström, P. (2018). Analytical sociology and computational social science. Journal of Computational Social Science, 1(1), 3-14.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. The quarterly journal of economics, 133(1), 237-293.

Lazer, D., & Radford, J. (2017). Data ex machina: introduction to big data. Annual Review of Sociology, 43, 19-39.

Merton, R. K., & Merton, R. C. (1968). Social theory and social structure. Simon and Schuster.

Molina, M., & Garip, F. (2019). Machine learning for sociology. Annual Review of Sociology.

Morgan, S. L., & Winship, C. (2015). Counterfactuals and causal inference. Cambridge University Press.

Page, S. E. (2015). What sociologists should know about complexity. Annual Review of Sociology, 41, 21-41.

Piketty, T., & Goldhammer, A. (2014). Capital in the Twenty-First Century. Harvard University Press. Retrieved November 5, 2020, from http://www.jstor.org/stable/j.ctt6wpqbc

Robles-Granda, P., Lin, S., Wu, X., D'Mello, S., Martinez, G. J., Saha, K., ... & Dey, A. D. (2020). Jointly Predicting Job Performance, Personality, Cognitive Ability, Affect, and Well-Being. arXiv preprint arXiv:2006.08364.

Sampson, R. J. (2012). Great American city: Chicago and the enduring neighborhood effect. University of Chicago Press.

Saavedra, M., & Twinam, T. (2020). A machine learning approach to improving occupational income scores. Explorations in Economic History, 75, 101304.

Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., ... & Datta, D. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. Proceedings of the National Academy of Sciences, 117(15), 8398-8403.

Sewell, W. H., Haller, A. O., & Ohlendorf, G. W. (1970). The educational and early occupational status attainment process: Replication and revision. American sociological review, 1014-1027.

Shirado, H., Iosifidis, G., Tassiulas, L., & Christakis, N. A. (2019). Resource sharing in technologically defined social networks. *Nature Communications*, *10*(1), 1079. https://doi.org/10.1038/s41467-019-08935-2

Suel, E., Polak, J. W., Bennett, J. E., & Ezzati, M. (2019). Measuring social, environmental and health inequalities using deep learning and street imagery. Scientific reports, 9(1), 1-10.

Toomet, O., Silm, S., Saluveer, E., Ahas, R., & Tammaru, T. (2015). Where do ethno-linguistic groups meet? How copresence during free-time is related to copresence at home and at work. PLoS One, 10(5), e0126093.

Watts, D. J. (2014). Common sense and sociological explanations. American Journal of Sociology, 120(2), 313-351.