A research program for dealing with most administrative data challenges: data linkage and latent variable modeling

Daniel L. Oberski

David Hand (DJH) has written an excellent overview of the challenges associated with using administrative data. As the paper already hints, many of the issues discussed apply equally to other types of "data exhaust", such as social media, internet usage data, and many other types of "big data". We must lay the groundwork of addressing these challenges before "data exhaust" can be leveraged by businesses for actionable insights and by scientists for valid conclusions. DJH's call to arms is therefore well-taken.

Here, I would like to take up that call by suggesting a single common theoretical framework for addressing what I consider the four most important challenges. The framework I suggest is:

1. Linkage of the "exhaust data" to *designed* data, followed by
2. Latent variable modeling (LVM).

While both surveys and administrative data, for instance, have errors and strengths, these errors are often non-overlapping, making the strengths complementary. For example, surveys directly address definition errors, and cover parts of the population missed by registers, while administrative data omit many errors associated with the survey response process. Of course, neither source is perfect in any respect. But linkage effectively "crosses" the errors, making it possible (identifiable) for LVMs to account for the impact of errors in both sources.

A recent example is our work on linked survey-administrative data to address **measurement error** (Oberski et al., 2017). Similarly, LVMs have been extended with graphical modeling of missing data to account for **selection bias**. Of course, the linkage itself may introduce errors as well. Such **linkage error** has been recognized as another form of LVM: a "latent class" or "finite mixture" model; including these classes in the overarching model mirrors the ideas in Lahiri and Larsen (2005) and is also suggested in Oberski et al. (2017). Finally, **privacy** is an important issue. The existing solutions to this issue – disclosure control, distributed computation, and homomorphic encryption – are all applicable to LVMs as well. Indeed, adding noise to hide values, as done in the disclosure control and differential privacy literatures, leads to an LVM with known measurement relationships.

Much more work is needed. But LVMs are a promising overarching framework to adress the challenges laid out so eloquently in this paper. Perhaps someday this framework could serve as the "generally accepted theory" that DJH identifies as currently lacking.

References

Lahiri, P. and Larsen, M. D. (2005). Regression analysis with linked data. *Journal of the American statistical association*, 100(469):222–230.

Oberski, D. L., Kirchner, A., Eckman, S., and Kreuter, F. (2017). Evaluating the quality of survey and administrative data with generalized multitrait-multimethod models. *Journal of the American Statistical Association*, (https://doi.org/10.1080/01621459.2017.1302338).