KfW Development Research
Development in Brief

New development research methods: machine learning

No 44, 1 December 2016

One Pager

Author: Dr Niels Kemper
Editor: Simone Sieler

In recent times, enormous methodological progress has been achieved in the scientific field of development economics. Historically, research in development economics has been characterised by theoretical models. Since the 1990s, empiricists have been dominating the field, who are using and refining regression models to analyse the factors of economic development (education, health, infrastructure etc.). The development of experimental methods (e.g. randomised control trials, RCTs) at the beginning of the new millennium meant a quantum leap, as the method enables researchers to estimate causal effects in a much better way than before. Today, research in development economics could face a new methodological quantum leap, which can be subsumed under the heading "machine learning". Related concepts are "neural networks" and "artificial intelligence".

How does machine learning work?

While researchers used to find it difficult to collect enough reliable data for their econometrical analyses, the volume of data today is often barely manageable ("big data"). This is driven by the rapid digitalisation of the world, the growth of the Internet, increasing computer capacities and decreasing storage costs for large volumes of data as well as by open data initiatives.

Very large data sets (big data) cannot be analysed with traditional regression analysis tools, even if these were developed for experimental data, as the number of variables typically exceeds the number of observation units (households, companies etc). Regression analysis is only useful when there is a small number of variables.

And this is where machine learning comes in: In contrast to previous methods, machine learning does not even try to estimate causal effects or describe them as precisely as possible. Instead, machine learning looks for "patterns" in a sheer endless number of variables and tries to differentiate these patterns in an iterative process. Machine learning assumes that the patterns found are the result of highly complex, previously unknown interdependencies. The patterns are assumed to be constant over time and can therefore be used to produce forecasts.

Small-scale weather forecasts are particularly suitable for this method. Atmospheric interrelationships are so complex that causal chains of effects are not enough to capture them. In the case of machine learning, the computer is fed instead with a huge number of historical data from weather stations. In order to make a forecast, scientists take current data from the weather station and look for similar constellations (patterns) in the past. They are thus able to make assumptions for the future based on past weather behaviour, which occurred after a similar starting situation. Each new set of weather data improves the quality of the forecasts with these models, i.e. the computer is constantly learning and producing better and better forecasts. This is why we speak of "machine learning".

What type of scientific problems can be addressed by machine learning?

Machine learning is a way to analyse highly complex and previously unknown interrelationships to produce probability prognoses regarding future developments. The key prerequisite is the availability of large amounts of data and of real patterns (i.e. the data is somehow related, in terms of content, to the projected values).

The method therefore seems generally fitting for development research: development processes are highly complex and it has not been possible in the past to describe them in their entirety with causal effects. In addition, data availability has increased dramatically over recent years, also in the developing countries. In addition to forecasts about economic development (value creation, trade flows etc.), these models could also be used to forecast SDG results (poverty, hunger, health etc.).

A widely recognised study in Rwanda, for instance, used anonymised mobile phone user data from the entire country (length and frequency of telephone conversations, mobility patterns etc.) in order to create regional poverty forecasts. These poverty forecasts now potentially serve the Rwandan government as a decision-making tool for the regional planning of promotional programmes.

Potential & limits for the application of this method in development research

Machine learning can be a meaningful tool for analysis and forecasts in development research. So far, not many machine learning applications are available for development cooperation contexts, but the more data is available, the more of these applications we will see. However, using machine learning processes also has its limits. The method can only be used in cases where historical data is available (i.e. results of innovative solutions cannot be forecast) and where relationships are constant over time (i.e. no structural breaks). For assessment and evaluation purposes, machine learning methods are not very suitable, as causal effects play a bigger role here than forecasts - and this is precisely what machine learning does not focus upon.■