Machine Learning as a data driven tool in result analysis

By Wolfgang Britz, Institute for Food and Resource Economics, University Bonn, Germany

Paper prepared for organized session:
„Result exploitation and analysis in large-scale economic models – state of the art and visions"

at the 15th Annual Conference on Global Economic Analysis

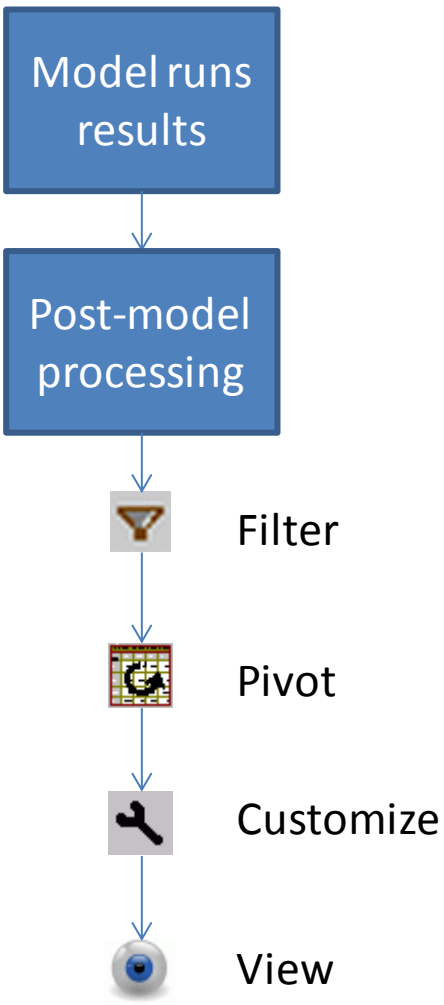
"New Challenges for Global Trade and Sustainable Development"

June 27-29, 2012, Geneva, Switzerland

Contents

Background

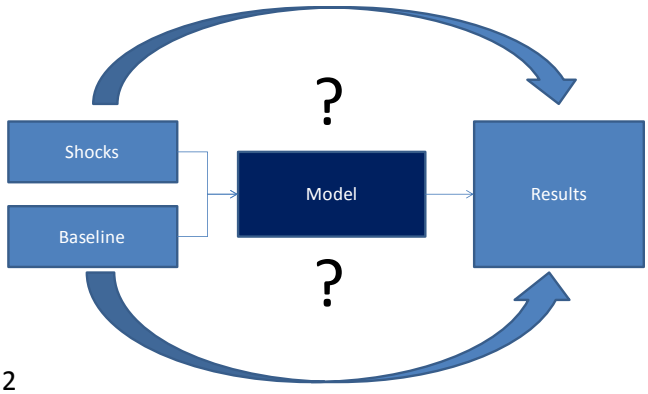
Huge amount of numerical data are generated by simulations with large-scale economic models generate, based on their high differentiation by sectors/products, in space or by policy instruments.



To give an example: a single counterfactual CAPRI run with the farm type layer switched on produces >20 Mio non-zeros. Lately, large-scale sensitivity analysis or using stochastic draws has come into fashion, increasing considerably the size of result sets. Besides the computational challenges related to managing large-scale data sets, the analyst is expected to dig up a story out of that haystack: What results are worth reporting? How can they be explained by cause-effect relations? What regions / sectors / institutions are winners or losers and why?

Since a long time, modelers have developed post model tools e.g. to derive indicators or to aggregate or decompose results as well as appropriate viewers. The graphic to the left depicts major steps in that process. Typically, from the vast amount of results produced, including indicators and decomposition results generated post-model, we filter out a smaller subset (e.g. only the GDP simulated by region). Next, we pivot the results in an appropriate manner (e.g. show the regions in the rows and the scenarios in the columns). We might then customize, e.g. by adding percentage changes against the baseline, choosing the font, adding units or explanatory long text. And finally, we show the result in an appropriate format, e.g. as a table, graphic or map. These approaches require a fair amount of priori knowledge and a priori decisions about what aspects to include in the analysis and how to present them.

Consequently, the resulting views are often already available pre-customized in a user interface such the user can simply select the view, possibly navigating through items. He thus benefits from past experience of other users who thought it worth to develop and store specific views of the model results. The tasks involved in visualizing model results are in detail discussed in Perez-Dominguez et al. 2012, questions relating to the indicator calculations and decompositions in McDougal and Britz 2012, two papers originally presented in the same organized session as the current one.



Especially if new features are added to a modeling system, such as the new environmental satellites in GTAP, or in case of large-scale sensitivity analysis, existing a priori knowledge might not be sufficient to define appropriate filters, indicators and decompositions approaches which capture main messages. The result set turns into a

haystack and the yet unknown core findings in the proverbial needle. The model than easily takes on a black box character: despite the fact that each single equation and variable might be clearly documented and understood, their interactions via complex feedback loops might prevent us from capturing clearly how the start point (= baseline), parameters and the shocks interact inside the model's structure such that a specific set of results emerges. Here, more data driven approaches might help.

The paper is organized as follows. A first section gives a short overview on CAPRI before the basics of machine learning are discussed. Next, the technical integration in the Graphical User Interface (GUI) of CAPRI is discussed, based on screenshots of an example, followed by a summary and conclusion. Throughout the paper, CAPRI, being of similar complexity as GTAP and other larger economic or integrated assessment models, only serves in here as an example how Machine Learning algorithms can be used and integrated in the work flow of analyzing results from large-scale models.

Overview on CAPRI

The CAPRI model (Britz and Witzke, 2011) is a global comparative-static partial equilibrium (PE) model with a strong focus on Europe, consisting of a supply and a market module. The supply module, covering the EU, Norway, Turkey and Western Balkans, comprises independent aggregate non-linear programming models representing approximately 50 crop and animal activities of all farmers at either regional (280 NUTS II sub-national regional units) or farm type level. The farm type level (Gocht and Britz, 2010) provides for the whole EU a consistent dis-aggregation from the regional level to about 1850 farm type models differentiated by farm specialization and economic size. Each programming model maximizes regional or farm type agricultural income at given prices with explicit consideration of the CAP support instruments, subject to technical constraints for feeding, young animal trade, fertilization, set-aside, a land supply curve and production quotas as well as possible further policy related constraints. The programming models cover in rich detail the different coupled and de-coupled subsidies of the so-called first Pillar 1 of the EU's Common Agricultural Policy (dealing with markets and income support), as well as major ones from the so-called Pillar 2 (dealing with measures classified as relating to Rural Development): Less Favoured Area support, agri-environmental measures, Natura 2000 support. The interaction between premium entitlements and eligible hectares for the so-called Single Farm Premium (SFP) of the CAP is explicitly considered, as are the different national SFP implementations, possibly remaining coupled payments for suckler cows and sheep & goat, and national so-called "specific support programs" under article 68 of Council Regulation (EC) No. 73/2009. Equally, support schemes for Norway are depicted in detail.

Prices for agricultural outputs are rendered endogenous based on sequential calibration (Britz 2008) between the supply models and a market module. The latter is a global spatial multi-commodity model covering 77 countries or country aggregates in 40 trade blocks and about 50 products. The Armington approach (Armington, 1969), assuming that the products are differentiated by origin, allows simulating bilateral trade flows and related bilateral as well as multilateral trade instruments, including tariff-rate quotas.

As usual in supply side and partial equilibrium approaches, the modules in CAPRI are written in physical quantities and try to capture policy instruments as close to the actual implementation as possible. Border protection, to give an example, comprises specific and/or ad valorem tariffs, possible under minimal import price regime which can be combined with bi-lateral or multi-lateral

open TRQs. The explicit consideration of policy instruments, physical quantities and a rich technology representation in combination with a high dis-aggregation in space eases the calculation of economic, social and environmental indicators. The latter comprise detailed GHG emissions, a Life-Cycle analysis of fossil energy use in agriculture and balances for Nitrogen, Phosphate and Potassium. In order to ease linkage to bio-physical modeling, a down-scaling component to 1x1 km grid cell can be linked to CAPRI (Leip et al. 2008). CAPRI has been intensively used for assessment of the different CAP reform steps (see e.g. Britz et al. 2012, Morredu 2011), but also to trade (e.g. Piketty et al. 2009) or environmental policy questions (e.g. Tukker et al. 2011, Pèrez Dominguez et al. 2009).

In opposite to GTAP (for a more detailed comparison between CAPRI and GTAP see Britz & Keeney 2010), CAPRI does not feature application dependent aggregation by sectors and regions, but is always applied (as most supply side model or PEs) in full dis-aggregation. Technically, it is realized in GAMS and steered by a Java based GUI (Britz 2011a).

Machine learning in CAPRI

Motivation

Two extensions in the last years triggered mainly the need to develop new solutions to result analysis in CAPRI: the spatial down-scaling component and the farm type layer. The first one breaks down major results the regional results based on statistical estimators to about two hundred thousand clusters of 1x1 km grid cells, it so far mostly resulted in improved mapping facilities. But is usage in policy relevant applications is so far limited, probably linked to the fact that we miss so far good approaches to summarize and analyze the tremendous amount of information generated by that component. The farm type layer as the second extension further dis-aggregates the sub-national administrative regional models to farm groups, increasing the number of supply models from ~280 to close to 2000. Results from these single models are then not only aggregated in space (to country and country block), but also by farm size and farm specialization. Due to the fact that many policy instruments of the Common Agricultural Policy are differentiated by farm characteristics, the farm-type layer is now rather standard in policy relevant applications. But exploiting the full richness of its results remains a challenge.

Both extensions in CAPRI can be seen as examples of the on-going expansion of modeling tools. The land use component in GTAP-AEZ (Hertel et al. 2009) provides a recent example for an extension of GTAP. Other modeling systems such as FAPRI (e.g. Westhoff et al. 2006) or AGLINK-COSIMO (ref) have in the last years started to develop stochastic baselines, providing another example of improvements generating large amount of results.

Due to the tremendous increase in the amounts of results produced by the new extensions of CAPRI, it was decided to explore the application of more data driven approaches to result analysis which do not require (strong) ex ante knowledge about cause and effect relations underlying the results. That step seemed a kind of logical progression in approaches supporting result analysis. The first step in that sequence consisted in offering pre-defined tables with major results respectively indicators. It guides user to start their analysis with results which have repeatedly been proven to be salient for policy impact analysis. In a second step, calculators and views were added which systematically decompose these results such that the user is steered towards a systematic analysis of differences in these major results which helps to develop a clear and sound story line. And now, the exploitation

tools of CAPRI are enriched by a package which allows for more flexible and data driven approaches to analysis results.

In order to do so, an existing Java based software package for Machine Learning was integrated in the CAPRI GUI in 2011. That extension is thus rather new and so far not widely used or explored. But given its potential, it seems worth to be discussed beyond the relatively small circle of CAPRI users. One could alternatively export key results and possible drivers in statistical packages and perform similar type of analysis, but it seemed inviting to explore a more seamless integration in the existing software architecture which at the same time offers new analytical approaches.

Why might it be inviting to rely on more data driven approaches? Tables/views which can be efficiently used by the human brain will typically comprise only a few items (e.g. UN 2009). A geographic choropleth map will typically even only show one single item as maps tend to have a higher dimensionality in the regional dimension. If the number of observations in one or several dimensions becomes large – e.g. by sensitivity analysis, by stochastic draws, by downscaling to the grid level – we need to condense the information.

The art in presenting results consists in finding a compromise between information amounts digestible by the human brain and preserving the necessary detail. Simply dropping observations or excluding items or only presenting summary statistics from the analysis carries clearly the risk to overlook important aspects or to oversimplify. If new modules / indicators are added to a tool or experiments comprise shocks never analyzed before, existing filters might miss the point and users might (not yet) possess the necessary knowledge to efficiently select the data needed to understand simulated outcomes. Imagine an even rather clear task such as the analysis of profit changes by farm type simulated by a counterfactual run with CAPRI. The first steps are standard: we might want to aggregate the individual farm type results by farm specialization/size and possible country. The results could be presented as a table. Next, the task is to explain why some cells in the farm typology are more affected more than others. What attributes (e.g. revenue shares, production intensity …) drive these differences? Given the amount of data and the richness of results, performing such analysis manually by navigating through tables or building new ones is tedious and might lead to path dependent results.

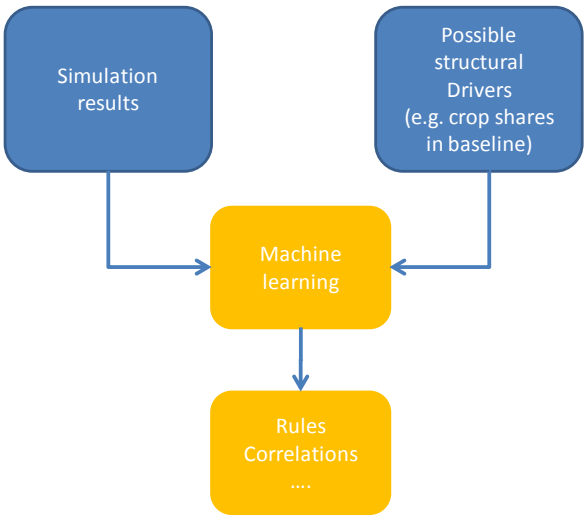
Computers are far better equipped to systematically analyze numerical relations in large-scale data sets, however, always with the risk to find spurious ones. Nevertheless, combining our knowledge about the model structure with the computing power of modern desktops might open up new avenues to result analysis. The data driven approach where algorithms try to “learn” from the data underlying patterns is called machine learning.

Application of machine learning for economic modeling result analysis

Definition of machine learning

Wikipedia gives the following definition: **“Machine learning**, a branch of artificial intelligence, is a scientific discipline concerned with the design and development of algorithms that allow computers to evolve behaviors based on empirical data, such as from sensor data or databases. Machine Learning is concerned with the development of algorithms allowing the machine to learn via inductive inference based on observation data that represent incomplete information about

statistical phenomenon. Classification which is also referred to as pattern recognition, is a important task in Machine Learning, by which machines "learn" to automatically recognize complex pattern, to distinguish between exemplars based on their different patterns, and to make intelligent decisions."

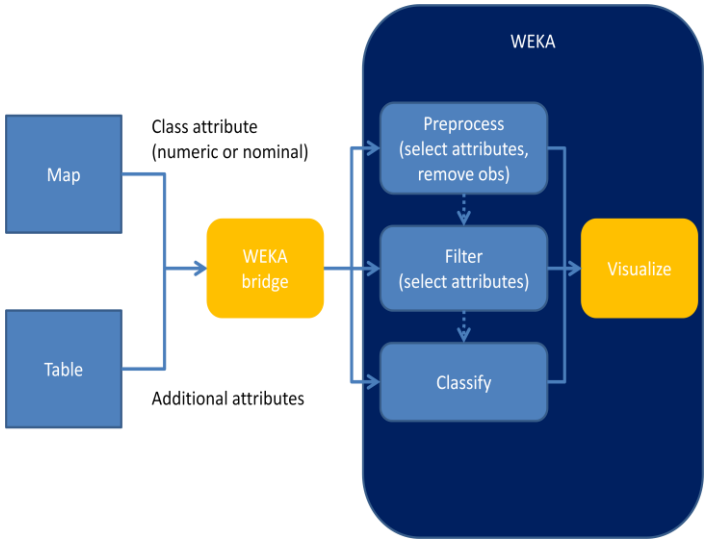


In the context of an economic model application, we might treat e.g. each region or farm type in our model as an "exemplar" taken from a data base. Some farm types might exhibit very large income changes, other little ones. What are common characteristics of the one and the other group? Machine learning might then come up with a "pattern" (e.g. based on a regression model) which determines the most important attributes impacting income changes in a given simulation. Machine learning has thus a lot of similarities with statistics – indeed many methods can also be found in statistical packages - but the focus to decide upon which attributes and relations matters is shifted to a certain extent from the human being to the computer. And, the tool box used in machine learning differs to a certain degree from classical statistics. Not at least, many of the algorithms had also been developed keeping computing time in mind.

One could also interpret many of the algorithms as "meta-modelling on the fly", i.e. to approximate complex quantitative cause-effects relations embedded in the equation structure and parameterization of our models into far simpler expressions. Originally, a simulation run with the economic model determines how a policy change affects our "exemplars", e.g. the individual regions or farm type. Machine learning might then derive from the results, ingoing data and parameters quantitative expressions which approximate the complex data generation process of the economic model. These far simpler expressions might then support the analyst in coming up with a clear story of how e.g. farm type attributes interact with changes in policy instruments to provoke the pattern of results we observe. If the expression e.g. indicates that the revenue share of cereals increases income in the simulation while the revenue share of ruminants has the opposite effect, we might be able to faster and with fewer errors to search for the underlying cause-effect relations or cluster the farm types or regions in our analysis accordingly.

Implementation in CAPRI

Since about a year, machine learning algorithms can be explored within the CAPRI GUI. The implementation in CAPRI is based on transparently integrating the WEKA machine learning library (Witten et.al. 2011) into the existing exploitations tool of the CAPRI GUI. The WEKA library as



an established tool is also integrated into other well known packages such as the data mining tool RapidMiner. Thanks to the GNU license which allows full access to the underlying Java source code, it was possible to more or less seamlessly link WEKA into the CAPRI exploitation tools which are also Java based and part of the CAPRI GUI (which also allows to steer the different model steps and perform other tasks such as linking remotely via internet to data and model repositories for updates). Only a few code changes were necessary to pass data transparently to the WEKA library from the tables and maps shown in the CAPRI GUI, without the need of further user interactions, underlining that that state-of-the-art object oriented programming approaches are used in WEKA.

The data exchange is done automatically in the background with the aim to reduce user input in the process. As a consequence, the powerful set of filtering, clustering and classification algorithm as well as related visualization tools from machine learning embedded in WEKA can be applied to the result sets inside the existing CAPRI exploitation tools (which comprise pre-defined views in form of tables, graphs and cloropeth maps plus functionality for customization by the user). The user does not need to first export data to another application to use the machine learning approaches, but can stay inside the same tools.

The current implementation is based on the **interaction of two views**:

1. A **map or table**– it defines the class attribute of the data

   The class attribute of the exemplar can be understood as the dependent variable of each observation. For classification algorithms which require nominal values, the assigned class from the classification underlying the color model of the map (which can be also applied to tabular views) is used. We might e.g. simply break the observations into quartiles and then let an algorithm predict to which quartile an observation belongs to.

2. A **table** with the “explanatory” attributes.

   These attributes are used to predict the class attributes. From that second table, the machine learning package is opened by a simple mouse-click.

The user can now either change the first view (e.g. by filtering the observations or choosing a different class attribute) or the second one to work on the explanatory attributes. In both cases, the updated numerical values are automatically passed to WEKA and the last algorithm chosen started again, so that updated results from the machine learning algorithm are immediately available.

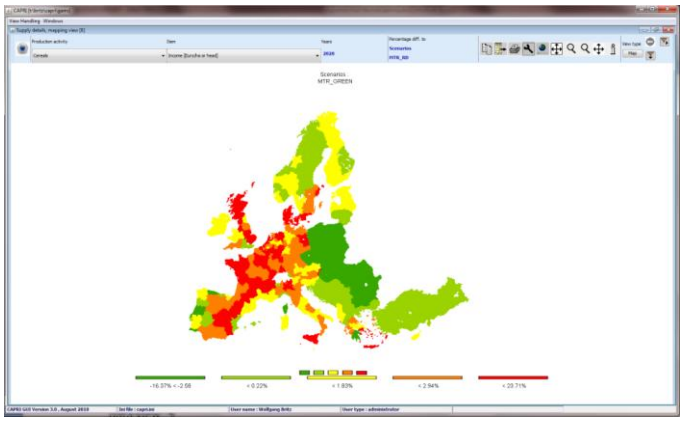
The WEKA library supports three different types of algorithms of immediate interest

1. **Filtering**: different algorithms which reduce the number of independent variables.
2. **Clustering**: algorithms which group the observations (the independent ones).
3. **Classification**: algorithms which “predict” the dependent variable. Here, besides methods known from statistics (such as multiple regressions) also algorithms which are not yet commonly used by economists such as trees or rules are offered.

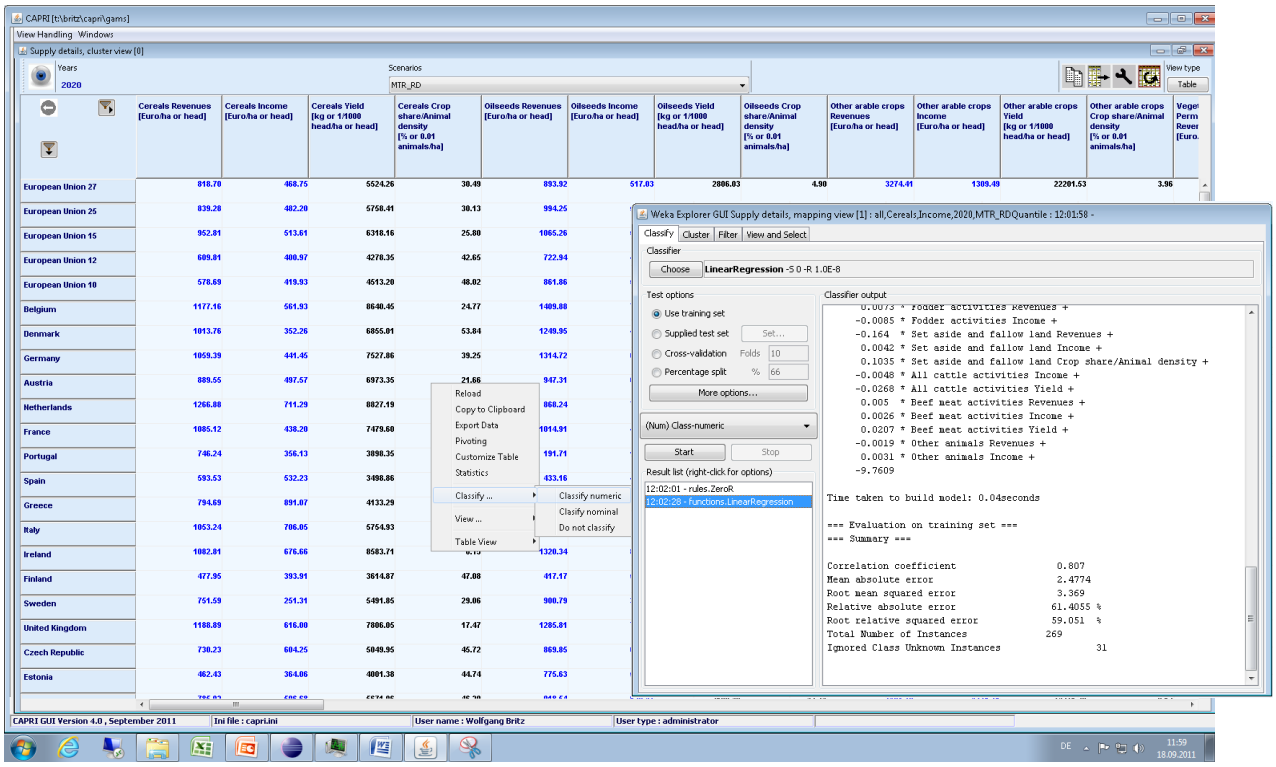
The basic data flow is shown in the graphic above.

Interaction between CAPRI GUI and WEKA

Let's construct an example: we want to check if the income change in cereals in a simulation depends on the crop shares of cereals and the yields. In order to do so, we first render a cloropeth map with the changes, which is a predefined view in CAPRI GUI to produce e.g. a map as shown on the right. The map as the first view hence delivers the dependent observations: either the individual numeric values or their nominal classes which determines the color used. Each region regions shown is hence an exemplar.



Graphic: Example for data to analyse



Graphic: Table with explanatory variables (in background) and view on the machine learning output.

The other attributes (independent variables) which are used for classification or clustering stem from a second table (see above). Technically, for the link to work, the observations from the map must be found in the rows of the second table. The user can then open the context menu of that table to chose "Classify" which then will open an instance of the WEKA graphical user interface as an additional individual window.

The classification is based on the complete functionality of the WEKA GUI regarding attribute selection/visualization, filtering and classification, see http://www.cs.waikato.ac.nz/~ml/index.html or http://www.capri-model.org/docs/WekaManual-3-6-5.pdf for a user manual. A typical output is

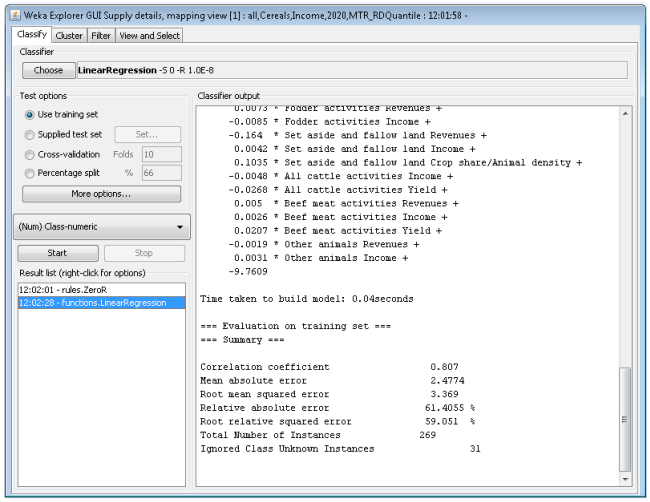
shown on top of the table with the explanatory attributes. In our example, we used as linear regression as the classifier, the output shows than the regression coefficients and further statistics.

The preliminary experiences with the machine learning package are quite positive. Firstly, in cases where the cause-effect relations are rather clear (at least to the experienced user) the classifiers are typically quite good in recover the drivers. Secondly, thanks to software implementation, using the package is quite easy and results are obtained rather fast.

The tabs “Classify”, “Cluster”, “Filter” and “View and select” allow the user to access specific part of the WEKA functionality. The result set from the current classification run can be shown in the lower left panel (result list). For each result set, a popup menu opens options, e.g. to show a graph with the prediction errors.

The WEKA classification panel

Classification algorithms predict from observed attributes (= independent, observed) the dependent ones. In Machine Learning is generally assumed that the independent attributes cannot be observed in the latter application, i.e. we aim to develop a good prediction. We could e.g. have an assembly line where an algorithm should classify the output into usable / not usable. We would first “train” the algorithm by telling him for so-called training instances if they are usable or not. The algorithm needs later to decide based on the observed attributes.



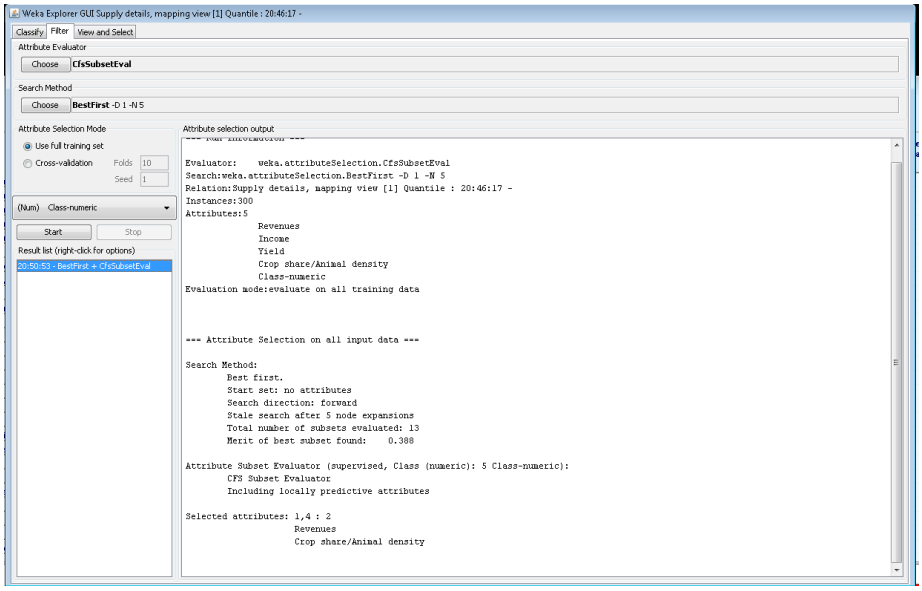
- The “**choose**” button will give access to a wide range of **different classifiers**, many of which have additionally options which can be edited by users. A multiple linear regression using the Akaide criterion for model selection is used as the default, assuming that most people will start with using numerical values as class attributes. Please note that switching between nominal and numerical class attributes might trigger error messages if the currently selected classifier cannot handle the newly selected class attribute type.
- It is recommended for our purposes to use under “Test options” “Use training set” (the default in our implementation) as we are typically not interested in an out-of-sample tests of the prediction quality.
- The actual classification can be started with the “start” button. If the data in the background are updated, the actually chosen classifier with the chosen options will be started on the new data set automatically. In absence of errors the “Classifier output” on the RHS will hence typically show results based on the latest selected data.

- The results can be visualized by clicking with the mouse on an item in the result list, the last on in the list always being the newest. If one has tried several classifiers, the old results remain available. However, if the data in the background change, the old results are automatically removed.

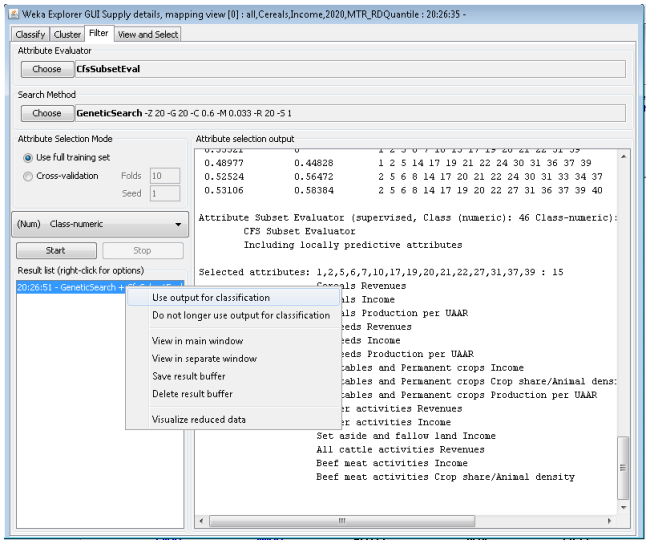
The reader should note that all the functionality described is identical to the standard WEKA GUI so that the user manual from WEKA can be used for further information.

PS: The cluster panel is not described, it works quite similar. Note however that filters are not applied to the cluster (see below).

The WEKA filtering panel



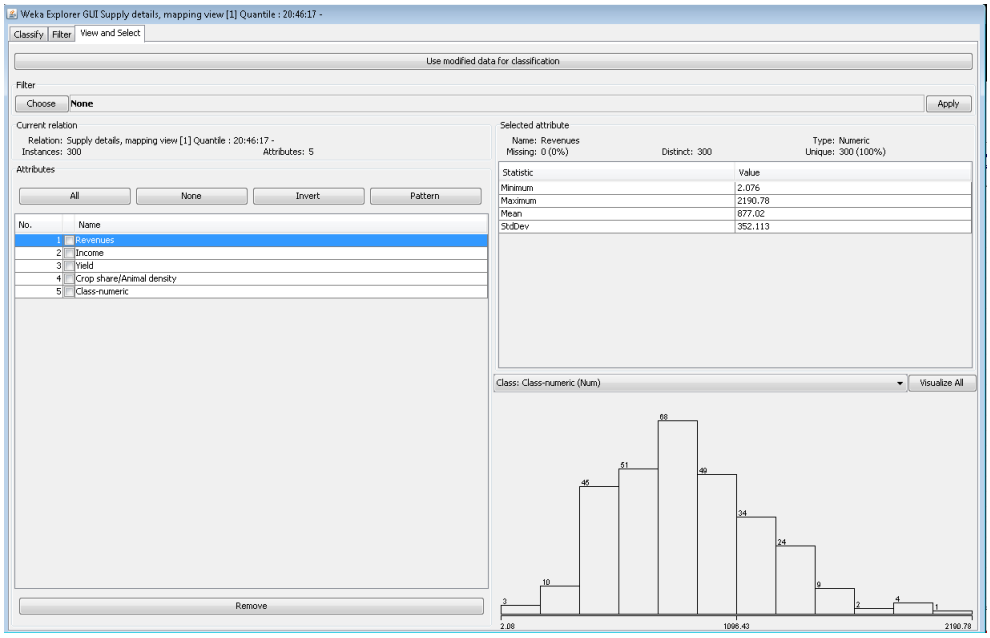
The filter panel allows running different types of filters which remove attributes, in many cases reflecting the correlation between attributes. In order to use the result from the filter run, click on the result set and chose “Use output for classification”:



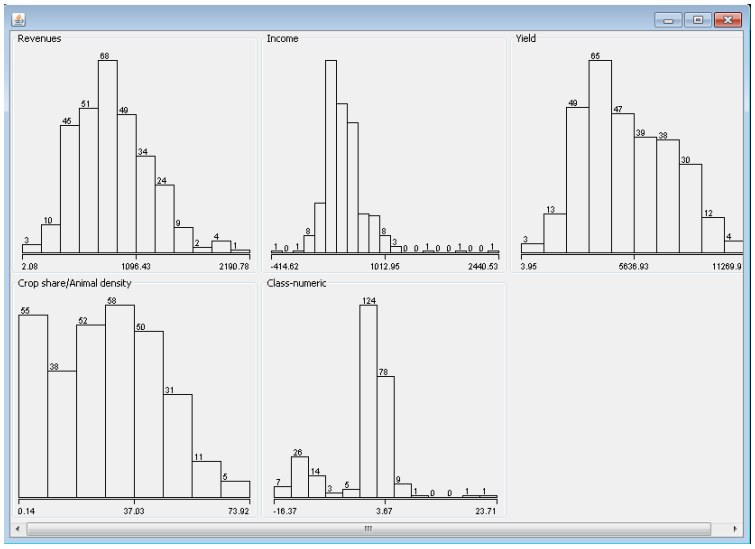
The last selected filter will be automatically restarted if a new data set is implicitly loaded (change of the map or of the data in the cluster table with the explanatory results). In order to switch off the use of the filter, select “Do not longer use output for classification”

The WEKA panel for attribute viewing and selection

The last panel available is especially interesting to quickly analyze statistics of the underlying data:



The reader can manually remove attributes and the reduced set of attributes will then passed to the filter and classifier. However, the attribute selection is not maintained when new data are loaded. The “Visualize All” button produces graphs of all current attributes:



Summary and conclusions

The analysis of results from large-scale economic models is an increasingly challenging field due to the development of new “satellites” combined with increasing scope and resolution of our tools. These developments reduces the number of generalist users who have still a full overview over the

different modules which might cover different domains (economic, social and environmental), might be embedded directly in the simulation model or work as post-model modules. Transparent and efficient approaches to analyze results are thus needed to countervail possible negative effects such as naïve interpretation of outcomes which might result from the growing specialization in model development and increased richness in model results.

The paper hinted at the potential benefit of using more data driven approaches which require less a priori knowledge while being flexible and fast enough to be applied interactively by the user. Such data driven approaches filter out automatically attributes which contribute little of explaining results of interest while characterizing relations between these attributes and the analyzed results. A multiple regression with automatic suppression of insignificant independent variables provides a well known example of such an approach.

The example of the integration of a Machine Learning package into the CAPRI GUI underlines the possible usefulness to extend our toolbox by data driven approaches. The actual software implementation proves that well developed packages can be linked in a transparent and elegant way in the tools we use to view and analyze results. The integration of the package into CAPRI is relatively new. As so often, early, more experimental adaptors will explore and then teach followers. It will hence take a while until we know if any of the algorithms will be more regularly applied.

References

Armington P (1969) A Theory of Demand for Products Differentiated by Place of Production, IMF Staff Papers 16, 159-76.

BRITZ W. (2008): *Automated model linkages: the example of CAPRI,* Agrarwirtschaft, Jahrgang 57 (2008), Heft 8

Britz W. (2011a). The Graphical User Interface for CAPRI version 2011, Institute for Food and Resource Economics, University Bonn (http://www.capri-model.org/docs/Gui2011.pdf).

Britz W. (2011b). Decomposing price and other effects in the CAPRI market model for result analysis. CAPRI technical note (http://www.capri-model.org/docs/Decomposing_market_model_results.pdf).

Britz W. and Keeney R. (2010): The CAPRI model - an overview with a focus on comparison to GTAP, selected paper presented at the Thirteenths Annual Conference on Global Economic Analysis "Trade for Sustainable and Inclusive Growth and Development", 9-11 June, 2010, Pennang (Malaysia)

Britz W. and Witzke P. (2008) CAPRI model documentation 2008: Version 2 (http://www.capri-model.org/docs/capri_documentation.pdf)

Britz W. and Witzke P. (2011) CAPRI model documentation 2011. http://www.capri-model.org/docs/capri_documentation.pdf

BRITZ, W., GOCHT, A., PÈREZ DOMINGUEZ, I., JANSSON, T., GROSCHE, S. AND ZHAO, N.: EU-Wide (Regional and Farm Level) Effects of Premium Decoupling and Harmonisation Following the Health Check Reform, German Journal of Agricultural Economics, vol. 61, p. 44-56

Gocht A. and Britz, W. (2011): EU-wide farm types supply in CAPRI - How to consistently disaggregate sector models into farm type models, *Journal of Policy Modelling*, 33(1): 146-167.

Hertel, T.W., S. Rose and R.S.J. Tol (2009). Land use in computable general equilibrium models: an overview. Chapter 1 in Economic analysis of land use in global climate change policy, T.W. Hertel, S. Rose and R.S.J. Tol (eds.). Routledge

Leip, A., Marchi, G., Koeble, R., Kempen, Britz W. and Li, C. (2008): Linking an economic model for European agriculture with a mechanistic model to estimate nitrogen losses from cropland soil in Europe, Biogeosciences, 5(1): 73-94

MORREDU, C. (2011) (MAIN AUTHOR AND EDITOR), WITH CONTRIBUTIONS BY MARTINI, R., KIMURA S., BRITZ, W., GOCHT, A., PEREZ, I., HART, K. AND BALDOCK, D.: Evaluation of Agricultural Policy Reforms in the European Union. OECD, Paris

Pèrez Dominguez I., Britz W. and Holm-Müller K. (2009): Trading Schemes for Greenhouse Gas Emissions from European Agriculture: A Comparative Analysis based on different Implementation Options, Review of Agricultural and Environmental Studies90(3), pp.287-308

Tukker A, de Koning A, Wolf O, Bausch-Goldbohm S, Verheijden M, Kleijn R, Pérez-Domínguez I and Rueda-Cantuche JM (2011): Environmental Impacts of Changes to Healthier Diets in Europe, Ecological Economics 70 (2011) 1776-1788, Elsevier

UN (2009): UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE, Making Data Meaningful, Part 2: A guide to presenting statistics, UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE, geneva, 59 pages, (http://www.unece.org/fileadmin/DAM/stats/documents/writing/MDM_Part2_English.pdf)

Westhoff, P.; Brown, S. Hart, C. (2006). When Point Estimates Miss the Point: Stochastic Modeling of WTO Restrictions. Journal of International Agricultural Trade and Development 2:87-107.

Witten I.H., Frank E., Hall M.A. (2011). Data Mining Practical Machine Learning Tools and Techniqes. Third edition. Elsevier, Amsterdam. 630 pages