Model building in Agricultural Economics with Machine Learning: Echoes from the past

C. Gardebroek[1,]* and M. Kornelis[2]

[1]Agricultural Economics and Rural Policy group, Wageningen University, Hollandseweg 1,
6716KN, Wageningen, The Netherlands
[2] Wageningen Social and Economic Research, Droevendaalsesteeg 4, 6708PB, Wageningen,
The Netherlands

Contributed Paper prepared for presentation at the 99th Annual Conference of the Agricultural Economics Society, Bordeaux School of Economics, University of Bordeaux, France

14 – 16 April 2025

*Agricultural Economics and Rural Policy group, Wageningen University, Hollandseweg 1, 6716KN, Wageningen, The Netherlands, koos.gardebroek@wur.nl

Abstract

Machine Learning tools are currently transforming empirical research in agricultural economics. However, a concern with these new tools is that they are purely data-driven. The history of economic science reveals a recurring tension between the roles of economic theory and data. The objective of this paper is to describe lessons from the apparent divergence between theory-driven and data-driven modelling approaches that can guide to the current rise of machine learning modelling in agricultural economics. We first discuss different views on using theory and data in economic building in general terms. Next, we review several key econometric papers in agricultural economics in order to show how economic theory and data are used. This is followed by an evaluation of agricultural economics publications that have employed machine learning techniques. Finally, we synthesize these findings in the discussion section and provide recommendations for the effective integration of machine learning in agricultural economics.

Keywords Machine Learning, econometric modeling, economic theory, data analysis

JEL code B41, C18, C51, Q00

1. Introduction

A new methodological wave in the form of Machine Learning tools currently freshens up empirical research in agricultural economics. Classic econometrics toolboxes are rapidly complemented with new data-driven algorithms that search for patterns in often large datasets to yield better forecasting or classification models (Storm et al., 2020; Brignoli et al., 2024). This data-driven approach makes economists wonder whether there is still a role for economic theory in model construction, and how these learning tools can be reconciled with well-established economic theory.

Interestingly, these concerns are not new. In economic research, two distinct quantitative approaches are commonly recognized: theory-driven and data-driven. Theory-driven models base their functional forms on behavioural theories, hypotheses, or a priori assumptions, which are then tested against real-world data. Conversely, data-driven models derive their structure from empirical observations, statistical tests, or data generalizations, from which theory is inferred. Both approaches have long been the subject of intense debate, with supporters on either side emphasizing their respective strengths and weaknesses. E.g. in the early 1980s a methodological debate arose with the arrival of data-driven time-series vector autoregressive (VAR) models that challenged existing structural econometric models that had been developed in the 1950-1970s, but that often failed to predict well in the 1970s (Juselius, 2009). Interestingly, this classic debate between 'theory first' and 'reality first' economic modelling approaches echoes in the current rise of machine learning in econometric model building.

Although this debate suggests a clear choice between theory-driven and data-driven methods, in practice, purely one-sided approaches are rare. Most real-world economic models blend elements from both approaches. For instance, Fraser and Moosa (2002) employed a Linearized Almost Ideal Demand System to study household food expenditures in Britain, modifying theory-based trend components with data-driven alternatives. Similarly, Gutierrez et al. (2014) used a Global Vector AutoRegression model to analyze global wheat export prices, replacing a data-driven long-term equilibrium with a theory-driven counterpart. The difference between a theory-driven or data-driven modelling approach is often whether the starting point lies in economic theory, or in data analysis.

The objective of this paper is to investigate how theory-driven and data-driven approaches have shaped (agricultural) economic research in the past, and what this implies for the current rise of machine learning modelling in agricultural economics. More precisely, this paper aims to answer the following research questions. First, how did theory-driven and reality-driven approaches originate and differ in general economics. Second, how have these two approaches influenced model building in agricultural economics. Third, what does contrast between theory-driven and data-driven model building imply for the current rise of machine learning in agricultural economic model building? In this paper we focus in particular on research involving time series data and econometrics since (i) this area was an earlier battleground between theory-driven and data-driven approaches, (ii) just like machine learning time series econometrics also seems less connected to economic theory, and (iii) both time series and ML have in common a strong focus on prediction..

The paper is structured as follows. Section 2 discusses historical development in theory-driven and data-driven econometric modelling in general. In section 3 we dive deeper in agricultural economic time series modelling in various publications. To what extent are their model components either theory- or data-driven and how have these classifications evolved over time. In section 4 we discuss the use of Machine Learning in general and assess how in recent publications in the agricultural economics literature using ML theory and data are used to construct these models. In the concluding section 5 we evaluate the potential impacts of machine-learning algorithms on future agricultural economic model development.

2. Theory-driven and data-driven approaches in past economic work

Economists have always differed in the importance attached to theory and/or data in their analyses. According to Hendry (2009) economists for a long time valued economic theory over empirical analysis of economic data. Theory was needed for a priori identification restrictions. This is confirmed by Spanos (2009) who states that theory-driven approaches represented the status quo in economics for a long time. The fact that for decades data was limitedly available also contributed to this pre-eminence of this theory-driven approach. And even if data was available, applied methods were often unreliable or had to deal with spurious correlations.

Gilbert and Qin (2005) discuss a variety of approaches in economic analysis in the pre-1940 period with mixed emphasis on data and theory. Some early studies were mainly data-driven to investigate economic phenomena such as business or pig cycles. Others were grounded in economic theory using data to parametrize their models. Often the observation was made that theory and empirics did not match perfectly. This led some researchers to conclude that existing theories were inadequate since they could not explain observed economic phenomena in the data (Gilbert and Qin, 2005). In other words, theory had to be adapted to become in line with the observed data. Others recognized that there were also problems and methodological issues with data, e.g. trends and spurious correlations, requiring a firm theoretical basis before any economic model should be quantified or interpreted.

A formalisation of econometric practices was established by the Cowles Commission in the 1940s and 1950s. Econometric analyses were grounded in a dynamic simultaneous equations system, based on Walrasian general equilibrium theory. Moreover, a solid body of econometric literature arose dealing with consistent estimation of parameters in such systems. That this framework was strongly theory-driven is illustrated by the fact that little attention was paid to diagnostic tests for model specification, since the underlying model was assumed correct anyway (Gilbert and Qin, 2005).

After WO2 a more consolidated body of econometric literature arose, with many classical textbooks as prime examples, e.g. Klein (1962), Johnston (1963), Goldberger (1964), and Malinvaud (1966), which mainly focused on quantifying economic relationships, recognizing the importance of economic theory. The availability of computers starting in the 1960s and econometric software in the 1970s further contributed to maturing the econometric field. Already in the 1970 computers were considered indispensable is doing large-scale analysis and dealing with non-linearities (Klein, 1971).

Gilbert and Qin (2005) describe how in the 1970s and 1980s econometrics developed in several directions, some strongly theory focused, others data-driven, and some trying to bridge theory-driven and data-driven approaches. Theory-driven econometric advancements were made in the areas of macroeconomics (rational expectations, dynamic models, growth theory) and the new field of micro-econometrics. Bayesian econometrics tried to bridge theory and data, by formulating the former as prior information to be combined with data (likelihood) leading to posterior distributions based on both theory and data.

A more data-driven approach in the form of multivariate time series models (VAR, VECM) became popular in the 1980s. Based on influential papers by Sargent and Sims (1977) and Sims (1980) the VAR approach provides a more data-driven approach to model identification. Instead of starting from a given dynamic model specification, Sims started to investigate dynamics from data to theory. In the decades that followed this data-driven approach expanded, mainly among European econometricians, whereas US econometrician sticked to more stylized theory based econometric models, e.g. dynamic stochastic general equilibrium models (DSGE) (Colander, 2009). Juselius (2009), who advocates this data-driven or 'reality first approach, attributes the popularity of theory-driven DSGE approaches as a response to the critique by

Summers (1991), who states that empirical economics has had little influence on economic theory development, which in itself already favors a theory-driven approach over a data-driven one. And even though in more recent times DSGE models have been expanded with e.g. VAR processes, Juselius (2009) still criticizes these approaches since they are not able to inform about changing parameters, competing theories, or new features in the data that have not been captured by the underlying theory. In fact, she clearly explains that in an inherent non-stationary world, starting with theoretical models that assume a stationary world, leads to wrong conclusions and ineffective policy advice, as witnessed during the 2007-2008 financial crisis. In the next section we investigate how theory and data are used to motivate model specification in time series modelling in agricultural economics.

3. Theory-driven and data-driven approaches in agricultural economics

This section reviews illustrative empirical applications of time-series models in agricultural economics, highlighting the distinction between theory-driven and data-driven approaches. Time-series econometrics has long been an area where this disconnect is particularly pronounced. This section follows a snowballing approach, beginning with empirical studies published in the American Journal of Agricultural Economics. We focus specifically on the early phase of the model-building process—before the model is confronted with data—where the divergence between theory-driven and data-driven methodologies becomes most evident. The objective is to assess the extent to which these empirical applications align with either approach in determining the model's functional form.

A time-series model comprises various components that collectively shape its functional structure and which can be configured in multiple ways. This study investigates whether these configurations can be categorized as theory-driven or data-driven. The components we examine include variable transformation, lags and leads, trends, cyclical fluctuations, sequential and contemporaneous causal relationships, long-term equilibria, structural shifts, and volatility clustering. These elements were selected based on our initial review of the American Journal of Agricultural Economics, where they appeared most frequently in empirical studies. Additionally, their inclusion allows for a comprehensive summary of the theory-driven versus data-driven approaches in time-series models in agricultural economics.

3.1. Variable transformation (taking logarithms) in the functional form

Concept. In economic time-series analysis, logarithms are often taken to capture the potential concavity of production, utility, and demand processes, to linearize potential exponential growth over time, to linearize multiplicative relationships, or to stabilize variance in the model, as economic time series often display increasing variance with magnitude (Nelson and Plosser 1982, p. 141). In all these situations, taking logarithms simplify the functional form's complexity.

Theory-driven illustrations. Thompson et al. (2002) studied the law of one price for wheat prices, for which the corresponding long-term equilibrium is a multiplicative relationship. Hauser and Andersen (1987), in their study about hedging with options under variance uncertainty, followed an option-pricing theory in which the log-price return is key. Liu and Shumway (2009) tested the induced innovation hypothesis for US agriculture, and modeled production constant elasticity of substitution (CES) function, which is nonlinear in its parameters. Chen et al. (2014) examined common forces that drive the prices of tradable commodities and followed previous studies who used logarithms. Ward and Davis (1978) also followed previous research in taking logarithms in their study about the effectiveness of coupon on consumption.

Data-driven illustrations. Gardner and Fullerton (1968) investigated the rental price of water, and used scatter diagrams of the underlying dataset to conclude to take logarithms of the variables.

3.2. Lags and leads in the functional form

Concept. The inclusion of lags can be justified by inertia—the tendency of economic behaviour to persist over time and resist immediate changes due to rigidities, habits, or institutional constraints. Conversely, the inclusion of leads in the functional form of a time-series model is often motivated by theoretical considerations such as expectations and forward-looking behaviour.

Theory-driven illustrations. Ward and Davis (1978) recognized the potential influence of habit persistency among consumers and, based upon existing theory, set up a model in which a lag of one period was included in the mathematical form of the model. Knudson (1991) studied the diffusion of semi-dwarf wheat innovations across the US, and his theoretical differential equations led to the inclusion of a single lag in the discretized time-series model.

Data-driven illustrations. Featherstone and Baker (1987) studied the impact of net-return and interest-rate changes on real farm asset values and applied a likelihood ratio test to determine the number of lags. Guney et al. (2019) investigated daily corn and soybean prices in North Carolina, and used Schwarz' Bayesian criterion to determine the number of lags. Liu and Shumway (2009) used Aikake's information Criterion, and Chen et al. (2014, p. 1458) used Ng and Perron (1995)'s general-to-specific rule to determine the number of lags.

3.3. Trending behaviour in the functional form

Concept. Time-series data often exhibit trending behavior over time, which can be classified as either deterministic or stochastic. A deterministic trend can be justified within economic models through growth theory, which attributes long-term economic expansion to fundamental real factors such as capital accumulation, population growth, and technological progress (Nelson and Plosser, 1982). These factors drive relatively smooth and gradual movements in economic activity.

In contrast, a stochastic trend follows a systematic pattern yet remains largely unpredictable (Maddala and Kim, 1998). This type of trend is often observed in economic variables like prices, which may follow a random walk—a concept closely tied to the efficient market hypothesis (Fama, 1965).

The presence of stochastic trends poses a challenge for many statistical methods, as they violate the assumption of stationarity required for valid inference (Dickey and Fuller, 1987). To mitigate this issue, differencing is commonly applied to remove stochastic trends, transforming the series into a stationary form by analyzing changes rather than absolute levels.

Theory-driven illustrations. To the best of our knowledge, we could not identify any AJAE publications that employ a theory-driven approach to operationalizing the inclusion of a stochastic trend in the functional form. Additionally, theoretical justifications for incorporating a deterministic time trend also appeared to be limited. The only illustration we found was Burt (1986), who examined factors that can explain farmland price fluctuations, and included a deterministic trend based on findings from a prior study.

Data-driven illustrations. Many studies can be found that use statistical tests for determining whether a deterministic or stochastic trend is appropriate. These studies rely on unit-root tests, which assess the presence of stochastic or deterministic trends in the underlying dataset (Dickey and Fuller, 1978). Examples include Ardeni (1989), who studied the law of one price for seven commodities across four countries; Goodwin and Holt (1999), who studied price transmission and asymmetric adjustment in the U.S. beef sector; Sephton (2003), who studied long-term relationships among corn and soybean prices in spatially separated markets; Balagtas and Holt (2009), who tested for the Prebisch-Singer hypothesis of a long-run decline in the relative prices of primary commodities; Ghimire and Griffin (2014), who studied water transfer effects of alternative irrigation institutions; and Ubilava (2017), who studied the consequences of El Niño Southern Oscillation (ENSO) shocks on primary commodity prices.

3.4. Recurring fluctuations in the functional form

Concept. Recurring fluctuations refer to periodic variations within a time series, such as seasonality and business cycles. Seasonality represents fixed, predictable patterns that occur within a defined timeframe, whereas business cycles capture broader economic fluctuations, including expansion and recession phases that can span several years. Seasonality can be deterministic, in which case shocks to the seasonal pattern dissipates over time, or stochastic, in which case shocks to the seasonal pattern have persistent effects.

Theory-driven illustrations. Rausser and Cargill (1970) examined the US broiler industry and used biological arguments to presume seasonal patterns. Hayenga and Hacklander (1970) used the existing pork storage fluctuations to theoretically justify seasonality in expected and live pork prices. Ying et al. (2019) followed previous research to assume seasonality in the future markets regarding US corn and soybean production. All these studies considered deterministic rather than stochastic seasonality.

Data-driven illustrations. Fraser and Moosa (2002), who considered a meat demand system in the UK, and Gómez et al. (2012), who studied food-inflation predictions in developing countries, both tested for stochastic against deterministic seasonality. Holt and Craig (2006) included seasonal dummies after visually observing seasonal patterns in their dataset of the hog-corn cycle.

3.5. Sequential causal relationships in the functional form

Concept. Sequential causal relationships between economic variables may exhibit one-way flows, where one variable influences other(s) without feedback in return, and feedback flows, where variables influence each other in both directions.

Theory-driven illustrations. Theory-driven operationalisations of sequential one-way flows imply an a priori classification of exogenous variables. For example Gutierrez et al. (2015) who studied the export prices of six regions, considered changes in oil prices and extreme weather events, as an a priori exogenous component in their model.

Data-driven illustrations. Orden and Fackler (1989) argue that theory often provides little guidance about lagged behavioural relationships, because past realizations of all variables are often known to economic agents and potentially will be used to form expectations. Data-driven operationalisations commonly start with a simultaneous equation model in which a priori all variable are treated as being endogenous (for example VAR or VECM frameworks), after which potential exogeneity is established through the use of statistical tests, such as Granger causality tests. Holt and Craig (2006) and Guney et al. (2019) belong to this stream of research.

3.6. Contemporaneous causal relationships in the functional form

Concept. Determining the direction of causality in a contemporaneous relationship is challenging, as the absence of an observable time lag complicates efforts to establish directly observable directionality. A common approach to addressing this issue is to establish a Cholesky decomposition of the covariance matrix. This method resolves current-period cross-correlations by imposing a causal ordering among the contemporaneous elements of the time-series variables under study. Bessler (1984a) emphasizes using economic theory for instantaneous correlations, considering biological lags in supply production, and accounting for market size and composition. Additionally, he highlights the role of market competition and structural aspects in selecting the appropriate autoregressive ordering.

Theory-driven illustrations. Bessler (1984b) investigated Brazilian agricultural prices, industrial prices, and money supply, herby using an a priori ordering of money supply, agricultural prices, and finally industrial prices. Similarly, Featherstone and Baker (1987) pre-imposed a one-way flow from interest rates to both returns and assets, as well as from returns to assets.

Data-driven illustrations. Holt and Craig (2006) and Guney et al. (2019) used the dataset-based variance-covariance matrix to estimate impulse responses (the so-called General Impulse Response function) to avoid imposing a variable ordering.

3.7. Long-term equilibria in the functional form

Concept. Arguments for the existence of long-term equilibrium relationships in economic systems include, among others, mean reverting in commodity cash prices (Schwarz, 1997), the law of one price (Isard, 1977), and the monetary neutrality hypothesis (Robertson and Orden, 1990). In contemporaneous time-series studies, the potential existence of a long-term equilibrium very often relates to the concept of cointegration (Engle and Granger, 1987). Cointegration occurs when two or more non-stationary time series respond similarly to market shocks and maintain a stable long-term relationship despite short-term fluctuations. The natural functional form to capture cointegration is the error-correction form.

Theory-driven illustrations. Goodwin et al. (2011), who studied North American oriented strand board markets, and Ardeni (1992) based their formulation of the long-term equilibrium upon the law of one price. Robertson and Orden (1990) used the monetary neutrality hypothesis to formulate a long-term relationship among the money supply, manufacturing prices and agricultural prices in New Zealand. Jin et al. (2012) built upon the formulations of the mean-reverting in commodity cash prices in their study of the South-Korean agricultural market. Hood and Dorfman (2015) based their long-term relationship upon the theory of spatial equilibria in prices and applied it in the US timber market. All these studies formulated the equilibrium in terms of a cointegrating relationship and therefore used an error-correction model.[1]

Data-driven illustrations. Adachi and Liu (2009) studied the Japanese pork market and based the formulation of the long-term equilibrium relationship on the outcome of break-point

[1] An example of a theory-driven application of an error-correction model that does not consider cointegration is Hallam and Zanoli (1992), who demonstrated how agricultural supply systems react to imbalances.

cointegration tests. Other examples include panel cointegration tests (Liu and Shumway, 2009), Johansen cointegration tests (Goodwin and Holt, 1999), and threshold cointegration tests (Sephton, 2003).

3.8. Structural transitions in the functional form

Concept. Agricultural markets can undergo significant events that reshape their long-term development. These may include technological innovations, policy shifts, capital flow disruptions, or demographic changes, all of which impact markets. In some instances such a transition process is smooth, whereas in others, the structural change can be abrupt (Maddala and Kim, 1998).

Theory-driven illustrations. Theory-driven operationalization for structural transitions can be found in Thompson et al. (2011), who investigated the structural impact of the CAP reform of 1992 on price transmissions elasticities in Germany, France, and the UK.

Data-driven illustrations. Goodwin et al. (2011) allow for structural transitions related to potential changes in production costs, demand, and non-competitive behaviour, and use information criteria and model fit statistics to choose among various structural transition functions. Wang and Tomek (2007) based the inclusion of a structural transition on visual inspection of their dataset on five commodity prices in Illinois. Enders and Holt (2012) used statistical tests to determine the number and timing of structural breaks in sixteen primary commodity prices.

3.9. Volatility clustering in the functional form

Concept. Actors in agricultural markets often experience fluctuations in market dynamics, where periods of high volatility are followed by periods of low volatility, and vice versa. This phenomenon, known as volatility clustering, reflects periods of heightened uncertainty, which can, in turn, influence decision-making among market participants.

A natural approach to studying volatility clustering involves the use of ARCH (Autoregressive Conditional Heteroskedasticity) and GARCH (Generalized Autoregressive Conditional Heteroskedasticity) models. ARCH models, introduced by Engle (1982), characterize conditional variance as a function of past squared errors. GARCH models, developed by Bollerslev (1986), extend this framework by incorporating past conditional variances, making them more effective for capturing long-term volatility dynamics.

Theory-driven illustrations. To the best of our knowledge, we have not found an empirical study that a priori superimposes the functional form of volatility clustering in a time-series model.

Data-driven illustrations. In their study on risk behaviour and rational expectations in the US broiler market, Aradhyula and Holt (1989), extended the rational expectations hypothesis to include price uncertainty. Jin and Frechette (2004), follow Baillie et al. (1996) who hypothesized that volatility persistence in agricultural futures prices is a consequence of staggered information flows. Hernandez et al. (2017) expect world market shocks to trigger volatility in the coffee prices of Ethiopian coffee. Ramirez et al. (2003) also consider heteroskedasticity in the futures contracts in the West Texas cotton basis, but give not argumentation why this phenomenon should occur. All these authors used statistical procedures based upon GARCH.

3.10. Key observations

Based on our review, we highlight several key observations:

- Certain components have consistently become data-driven across all the studies examined. In state-of-the-art applications, decisions regarding lag length, trending behavior, and volatility clustering are primarily determined by data.
- The theoretical foundations of the examined model components remain evident. In most cases, these theoretical underpinnings are explicitly stated to justify the inclusion of a particular component in the model's functional form, even if the subsequent operationalization was data-driven.
- When direct observation of a component is challenging, theory-driven approaches are often employed to define its functional form. Examples include contemporaneous relationships and long-term equilibrium relationships.
- It is crucial to distinguish between concept and operationalization. In most cases, a component's conceptual basis must first be theoretically justified before its inclusion in the functional form. Once justified, its operationalization may follow either a theory-driven or data-driven approach, depending on the researcher's choices. For instance, Bessler (1990) identified a potential seasonal pattern in an unspecified dataset but did not include a seasonality component in his model, as the concept lacked a solid theoretical justification.
- Many empirical studies incorporate both theory-driven and data-driven approaches within the same framework. A notable example is Ardeni (1992).
- It is also possible for the same component to be operationalized using both approaches. For example, Jin and Kim (2012) identified a structural break in 1995 in South Korea's agricultural market opening based on previous research—a theory-driven approach. However, in their empirical analysis, they also allowed for multiple additional structural breaks before and after 1995, a data-driven approach.
- Many studies become data-driven because they evaluate multiple alternative models and ultimately select the one that best fits the data. An example of this approach is found in Fraser and Moosa (2002).

4. Machine learning modeling in agricultural economics

Machine learning techniques are increasingly used in agricultural economics (see e.g., Storm et al., 2020 or Bayliss et al., 2021). An interesting question is how theory and data are connected in this recent literature, and whether it differs from past empirical agricultural economic studies. Before looking at machine learning applications in more detail, first some general notions about machine learning can be made, which are relevant for the comparison between theory-driven and data-driven approaches.

First, the main objective of machine learning algorithms is prediction, whereas more traditional econometric approaches also allow for estimation and attribution (Efron, 2020). The latter e.g., relate to estimating relevant policy variables, such as elasticities, or assessing whether and by how much certain variables contribute to outcomes of interest. Estimation and attribution also allow for hypothesis testing, in other words testing theory. Pure prediction algorithms are only able to assess whether certain predictions are in line with economic theory, without shedding much light regarding the uncertainty of these predictions, let alone concluding which variables are driving these outcomes. In short, the pure focus of machine learning methods on prediction make them less suitable to relate to economic theory, and are more inviting for a data-driven approach.

Second, a strength of machine learning is that it often is able to select subsets of variables from big datasets that provide the best cross-validated test (prediction) fit, e.g. LASSO or Elastic Nets. In other words, it is not economic theory that prescribes the choice of variables, but these variables are chosen in a pure data-driven approach. Moreover, several machine learning methods use ensembles of different models such as random forests, with different predictors included in submodels. In other words, there is a not a distinct set of variables which could provide theoretical insights and variables may vary from case to case.

Third, an often acclaimed strength of machine learning prediction models is that they are well able to capture non-linearities and structural breaks in the data. However, economic theory is often described in linear mathematically tractable models. Although non-linearities and breaks may be known features of real economic data, the underlying models often do not allow for them, except for corner solutions and threshold models.

4.1 The role of theory in recent agricultural economic studies using machine learning

To be added

5. Conclusions

Given this paper's objective to investigate how theory-driven and data-driven approaches have shaped (agricultural) economic research and what this implies for the use of machine learning in agricultural economics, three research questions were formulated: (i) how did theory-driven and reality-driven approaches originate and differ in general economics; (ii) how have these two approaches influenced model building in agricultural economics; (iii) what does the contrast between theory-driven and data-driven model building imply for the current rise of machine learning in agricultural economic model building?

Even though in economics some approaches are more theory-driven, whereas others suggest a more data-driven approach, it can be concluded that there was always a connect between theory and data, despite the different emphases. One could say that in data-driven research the theory is more open to debate, whereas in theory-driven research data limitations are often stressed. In agricultural economics studies this connect between theory and data is present just as well. Although specification of certain model components sometimes seem purely data-driven, many studies do refer to economic theory in order to justify the model as whole.

So, what does this imply for the growing body of literature using ML that can be expected in the coming years? It is important to realize that the strength of ML is prediction, something it has in common with time-series econometrics. However, its inability for testing hypotheses makes it less suitable for testing theories. In that sense there seems to be less connection to theory. Theory could of course still be used in various model components, such as functional form, variable selection, cointegration, etc. but also here the strengths of ML seem to make economic theory less needed. ML shines in automated variable selection, approximating unknown non-linear functional forms, and of course dealing with large sets of variables. All in all this could redefine the connection between economic theory and modelling, with outcomes of data-driven models more often proposing suggestions for revisions of theory, instead of the given status of theory informing current model building.

References

Aradhyula, S. V., & Holt, M. T. (1989). Risk behavior and rational expectations in the U.S. broiler market. American Journal of Agricultural Economics, 71(4), 892-902.

Ardeni, P. G. (1989). Does the law of one price really hold for commodity prices? American Journal of Agricultural Economics, 71(3), 661-669.

Baillie, R. T., Bollerslev, T., & Mikkelsen, H. O. (1996). Fractionally integrated generalized autoregressive conditional heteroskedasticity. Journal of Econometrics, 74(1), 3–30.

Baylis, K., Heckelei, T., & Storm, H. (2021). Machine learning in agricultural economics. In *Handbook of Agricultural Economics* (Vol. 5, pp. 4551-4612). Elsevier.

Bessler, D. A. (1984a). An analysis of dynamic economic relationships: an application to the US hog market. Canadian Journal of Agricultural Economics/Revue canadienne d'agroeconomie, 32(1), 109-124.

Bessler, D. A. (1984b). Relative prices and money: a vector autoregression on Brazilian data. American Journal of Agricultural Economics, 66(1), 25-30.

Bessler, D. A. (1990). Forecasting multiple time series with little prior information. American Journal of Agricultural Economics, 72(3), 788-792.

Bieri, J., & Schmitz, A. (1970). Time series modeling of economic phenomena. American Journal of Agricultural Economics, 52(5), 805-813.

Bjornson, B. (1994). Asset pricing theory and the predictable variation in agricultural asset returns. American Journal of Agricultural Economics, 76(3), 454-464.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics, 31(3), 307-327.

Brignoli, P. L., Varacca, A., Gardebroek, C., & Sckokai, P. (2024). Machine learning to predict grains futures prices. *Agricultural Economics*. Early Access.

Burt, O. R. (1986). Econometric modeling of the capitalization formula for farmland prices. American Journal of Agricultural Economics, 68(1), 10-26.

Chambers, R. G. (1979). Review of Forecasting Economic Time Series by Granger and Newbold. American Journal of Agricultural Economics, 61(3), 582-583.

Chavas, J.-P., & Holt, M. T. (1991). On nonlinear dynamics: The case of the pork cycle. American Journal of Agricultural Economics, 73(3), 819-828.

Chen, S.-L., Jackson, J. D., Kim, H., & Resiandini, P. (2014). What drives commodity prices? American Journal of Agricultural Economics, 96(5), 1455-1468.

Clark, J. S., Fulton, M., & Scott, J. T., Jr. (1993). The inconsistency of land values, land rents, and capitalization formulas. American Journal of Agricultural Economics, 75(1), 147-155.

Clark, J. S., & Youngblood, C. E. (1992). Estimating duality models with biased technical change: A time series approach. American Journal of Agricultural Economics, 74(2), 353-360.

Colander, D. (2009). Economists, incentives, judgment, and the European CVAR approach to macroeconometrics. *Economics*, 3(1), 20090009.

Criddle, K. R., & Havenner, A. M. (1989). Forecasting halibut biomass using system theoretic time-series methods. American Journal of Agricultural Economics, 71(2), 422-431.

Criddle, K. R., & Havenner, A. M. (1990). Forecasts from a state space multivariate time-series model. American Journal of Agricultural Economics, 72(3), 793-798.

Dorfman, J. H. (1993). Bayesian efficiency tests for commodity futures markets. American Journal of Agricultural Economics, 75(5), 1206-1210.

Dorfman, J. H., & McIntosh, C. S. (1991). Results of a price-forecasting competition: Reply. American Journal of Agricultural Economics, 73(4), 1277-1278.

Efron, B. (2020). Prediction, Estimation, and Attribution. Journal of the American Statistical Association, 115(530), 636-655.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica, 50(4), 987-1007.

Engle, R. F., & Granger, C. W. J. (1987). Cointegration and error correction: Representation, estimation, and testing. Econometrica, 55(2), 251–276.

Falk, B. (1991). Formally testing the present value model of farmland prices. American Journal of Agricultural Economics, 73(1), 1-10.

Featherstone, A. M., & Baker, T. G. (1987). An examination of farm sector real asset dynamics: 1910-85. American Journal of Agricultural Economics, 69(3), 532-546.

Ferris, J. N. (1974). Use of ratios and gross margins in time series supply analysis. American Journal of Agricultural Economics, 56(5), 1199-1212.

Flacco, P. R., & Larson, D. M. (1992). Nonparametric measures of scale and technical change for competitive firms under uncertainty. American Journal of Agricultural Economics, 74(1), 173-176.

Fraser, I., & Moosa, I. A. (2002). Demand estimation in the presence of stochastic trend and seasonality: the case of meat demand in the United Kingdom. *American Journal of Agricultural Economics*, *84*(1), 83-89.

Gilbert, C. L., & Qin, D. (2005). The first fifty years of modern econometrics (No. 544). *Working Paper.*

Goldberger, A.S. (1964). *Econometric Theory.* New York: Wiley.

Goodwin, B. K., & Holt, M. T. (1999). Price transmission and asymmetric adjustment in the U.S. beef sector. American Journal of Agricultural Economics, 81(3), 630-637.

Gould, B. W., Cox, T. L., & Perali, F. (1991). Demand for food fats and oils: The role of demographic variables and government donations. American Journal of Agricultural Economics, 73(1), 212-221.

Guney, S., Goodwin, B. K., & Riquelme, A. (2019). Semi-parametric generalized additive vector autoregressive models of spatial basis dynamics. American Journal of Agricultural Economics, 101(2), 541-562.

Gutierrez, L., Piras, F., & Paolo Roggero, P. (2015). A global vector autoregression model for the analysis of wheat export prices. *American journal of agricultural economics*, *97*(5), 1494-1511.

Hallam, D., & Zanoli, R. (1993). Error correction models and agricultural supply response. European Review of Agricultural Economics, 20(2), 151-166.

Hauser, R. J., & Andersen, D. K. (1987). Hedging with options under variance uncertainty: An illustration of pricing new-crop soybeans. American Journal of Agricultural Economics, 69(1), 38-45.

Hayenga, M. L., & Hacklander, D. (1970). Monthly supply-demand relationships for fed cattle and hogs. American Journal of Agricultural Economics, 52(4), 535-544.

Hernandez, M. A., Rashid, S., Lemma, S., & Kuma, T. (2017). Market institutions and price relationships: The case of coffee in the Ethiopian commodity exchange. American Journal of Agricultural Economics, 99(3), 683-704.

Hetemäki, L., & Kuuluvainen, J. (1992). Incorporating data and theory in roundwood supply and demand estimation. American Journal of Agricultural Economics, 74(4), 1010-1018.

Holmes, R. A. (1968). Combining cross-section and time-series information on demand relationships for substitute goods. American Journal of Agricultural Economics, 50(1), 56-65.

Jappelli, T., & Pistaferri, L. (2010). Title of the study. American Economic Review, 100(4), 2050-2070.

Jin, N., Lence, S., Hart, C., & Hayes, D. (2012). The long-term structure of commodity futures. American Journal of Agricultural Economics, 94(3), 718-735.

Johnston, J. (1963). *Econometric Methods.* New York: McGraw-Hill.

Juselius, K. (2009). Special issue on using econometrics for assessing economic models—An introduction. *Economics*, 3(1), 20090028.

Klein, L.R. (1962). *An Introduction to Econometrics.* Englewood Cliffs: Prentice Hall.

Klein, L. R. (1971). Whither econometrics? *Journal of the American Statistical Association*, 66(334), 415-421.

Knudson, M. K. (1991). Incorporating technological change in diffusion models. American Journal of Agricultural Economics, 73(3), 724-733.

Malinvaud, E. (1966). *Statistical Methods of Econometrics*. Studies in Mathematical and Managerial Economics, Vol. 6. Amsterdam: North-Holland.

McNulty, M. S., & Huffman, W. E. (1992). Trading-day variation: Theory and implications for monthly meat demand. American Journal of Agricultural Economics, 74(4), 1003-1009.

Maddala, G. S., & Kim, I. (1998). Unit roots, cointegration, and structural change. Cambridge University Press.

Moss, C. B., & Shonkwiler, J. S. (1993). Estimating yield distributions with a stochastic trend and nonnormal errors. American Journal of Agricultural Economics, 75(4), 1056-1062.

Mount, T. D. (1989). Policy analysis with time-series econometric models: Discussion. American Journal of Agricultural Economics, 71(2), 507-508.

Orden, D., & Fackler, P. L. (1989). Identifying monetary impacts on agricultural prices in VAR models. American Journal of Agricultural Economics, 71(2), 495-502.

Parvin, D. W. (1973). Estimation of irrigation response from time-series data on nonirrigated crops. American Journal of Agricultural Economics, 55(1), 73-76.

Ramírez, O. A., Misra, S. K., & Nelson, J. (2003). Efficient estimation of agricultural time series models with nonnormal dependent variables. American Journal of Agricultural Economics, 85(4), 1029-1040.

Rausser, G. C., & Cargill, T. F. (1970). The existence of broiler cycles: An application of spectral analysis. American Journal of Agricultural Economics, 52(1), 109-121.

Robertson, J. C., & Orden, D. (1990). Monetary impacts on prices in the short and long run: Some evidence from New Zealand. American Journal of Agricultural Economics, 72(1), 160-171.

Rozelle, S., & Boisvert, R. N. (1993). Grain policy in Chinese villages: Yield response to pricing, procurement, and loan policies. American Journal of Agricultural Economics, 75(2), 339-349.

Sargent, T. J., & Sims, C. A. (1977). Business cycle modeling without pretending to have too much a priori economic theory. *New methods in business cycle research*, 1, 145-168.

Schmitz, A., & Watts, D. G. (1970). Forecasting wheat yields: An application of parametric time series modeling. American Journal of Agricultural Economics, 52(2), 247-254.

Schroeter, J., & Azzam, A. (1991). Marketing margins, market power, and price uncertainty. American Journal of Agricultural Economics, 73(4), 990-999.

Schwartz, E. S. (1997). The stochastic behavior of commodity prices: Implications for valuation and hedging. The Journal of finance, 52(3), 923-973.

Sephton, P. S. (2003). Spatial market arbitrage and threshold cointegration. American Journal of Agricultural Economics, 85(4), 1041-1046.

Sexton, R. J., Kling, C. L., & Carman, H. F. (1991). Market integration, efficiency of arbitrage, and imperfect competition: Methodology and application to U.S. celery. American Journal of Agricultural Economics, 73(3), 568-580.

Sheldon, I. (1990). Review of Market Response Models: Econometric and Time Series Analysis by Hanssens, Parsons, & Schultz. American Journal of Agricultural Economics, 72(4), 1099-1100.

Shui, S., Beghin, J. C., & Wohlgenant, M. (1993). The impact of technical change, scale effects, and forward ordering on U.S. fiber demands. American Journal of Agricultural Economics, 75(3), 632-641.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, 1-48.

Skoufias, E. (1993). Seasonal labor utilization in agriculture: Theory and evidence from agrarian households in India. American Journal of Agricultural Economics, 75(1), 20-32.

Storm, H., Baylis, K., & Heckelei, T. (2020). Machine learning in agricultural and applied economics. *European Review of Agricultural Economics*, 47(3), 849-892.

Taylor, C. R. (1973). A production function model for aggregate time-series data. American Journal of Agricultural Economics, 57(1), 122-123.

Thompson, R. P., Capps, O. Jr., & Massey, J. G. (1994). Demand for an undergraduate education in the agricultural sciences. American Journal of Agricultural Economics, 76(2), 303-312.

Thompson, S. R., Sul, D., & Bohl, M. T. (2002). Spatial market efficiency and policy regime change: Seemingly unrelated error correction model estimation. American Journal of Agricultural Economics, 84(4), 1042-1053.

Todd, R. M. (1989). Policy analysis with time-series econometric models: Discussion. American Journal of Agricultural Economics, 71(2), 509-510.

Vukina, T., & Anderson, J. L. (1993). A state-space forecasting approach to optimal intertemporal cross-hedging. American Journal of Agricultural Economics, 75(2), 416-424.

Ward, C. E. (1992). Inter-firm differences in fed cattle prices in the southern plains. American Journal of Agricultural Economics, 74(2), 480-485.

Ward, R. W., & Davis, J. E. (1978). A pooled cross-section time series model of coupon promotions. American Journal of Agricultural Economics, 60(3), 393-401.

Whittaker, J. K., & Bancroft, R. L. (1979). Corn acreage response-function estimation with pooled time-series and cross-sectional data. American Journal of Agricultural Economics, 61(3), 551-553.

Ying, J., Chen, Y., & Dorfman, J. H. (2019). Flexible tests for USDA report announcement effects in futures markets. American Journal of Agricultural Economics, 101(4), 1228-1246.