DOI: 10.1111/joes.12505

ARTICLE

JOURNAL OF ECONOMIC SURVEYS WILEY

Measuring stability and structural breaks: Applications in social sciences

Daria Loginova | **Stefan Mann**

Department Socioeconomics, Research Division Competitiveness and System Evaluation, Research Station Agroscope, Federal Office of Economics, Ettenhausen, Switzerland

Correspondence
Daria Loginova, Department Socioeconomics, Research Division Competitiveness and System Evaluation, Research Station Agroscope, Tanikon 1, Ettenhausen, 8356, Switzerland.
Email: daria.loginova@agroscope.admin.ch

[Correction added on 07 May, 2022 after first online publication: CSAL funding statement has been added.]

Abstract

Several theoretical and methodological works have helped to clarify a number of principles of data analysis. The application of this knowledge in social studies when investigating stability and structural breaks needs a clear roadmap. This article suggests methodological frameworks that use structured knowledge on volatility and stability. It explains principles of measure selection per se and regarding research interest and relates these principles to statistical methods. The paper makes recommendations for practitioners concerned with variation issues on model and method selection.

KEYWORDS
measure selection, model selection, stability, volatility

1 | INTRODUCTION

Stability and structural breaks can occur during interventions and regulations, changes in formulations and climate, after treatment, between different social and economic groups, between homogenous and heterogeneous objects, they can be planned or natural and they may affect social and economic indicators. Stability and structural breaks are, therefore, a concern in most research on social and economic processes. Despite the growing political interest, the amount of data available and the increasing number of interventions, the roadmaps for measuring stability in many policy areas and social sciences have found few attention in literature compared to software,

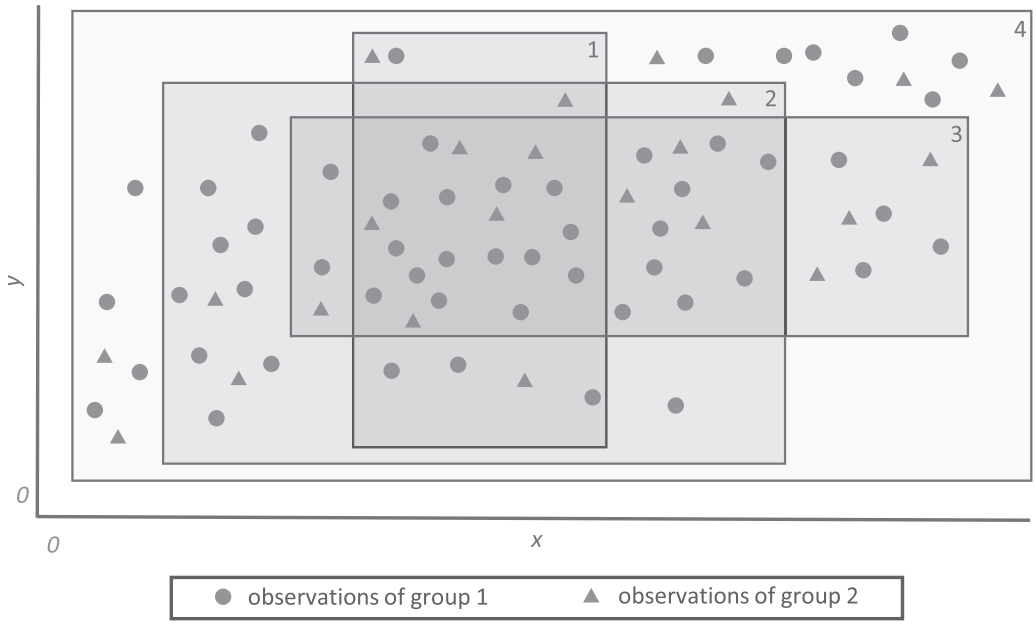

FIGURE 1 *Visualisation of a data pattern*. Note. Each square collects the observations of interest. The number of a square is denoted in top right apex

modeling and computational issues in data analysis. This paper fills this gap and focuses on various issues related to stability measurement and modeling.

There are excellent works that have summarized the methods for studying data (e.g., Athey & Imbens, 2019; Greene, 2011; Lantz, 2019; Wooldridge, 2013). The number of options for data investigation is enormous, and there is no single uniform methodology for each single research question (e.g., Athey & Imbens, 2019; Maggino & Facioni, 2017). Let Figure 1 visualize the data template. In measuring stability or structural break, the definitions (the choice of x and y in a graph), measure selection (the scales of x and y) and model selection (types of the dots and the selection of the subsets 1–4) decisively influence the results of a study.

This article summarizes the basic knowledge and techniques on volatility and stability measurements in social sciences, explains their advantages and disadvantages, and suggests several ways of volatility and stability investigation under different circumstances. In contrast to other methodological contributions, we look at statistics from the perspective of a specific research task—the study of stability and structural break in various understandings of these terms.

We focus mainly on the idea behind the methods and measures, comparing them, but removing the mathematical descriptions. Our general goal was to address the practical issues of a wide range of measures and methods, and to guide the readership's choice between them. We draw readers' attention to potential practical problems in the application of the tools in social sciences and suggest possible solutions. Finally, we provide a roadmap for data investigation that helps navigating between many methods, avoiding pitfalls and applying model modifications.

For this purpose, we first introduce some basic understanding of volatility and stability in Section 2 and then consider principles of volatility measure selection in Section 3. After, we present the background information on the methods in Section 4 and suggest measurement options and

statistical methods following the possible state of research interest in Section 5. Section 6 concludes the study.

2 | DEFINITIONS

2.1 | Development

The first important definition is the understanding of "development." Although in different research areas the term "development" is understood differently (see general perspective in Cambridge dictionary; in socioeconomics: Myrdal, 1974; Rabie, 2016; in land development: Johnson, 2008), to make the "development" measurable, any of them should consider at least the following five main cornerstones: (1) the investigated object; (2) the criteria and the rules for tracking the changes in this object; (3) the efforts, influences, environment and conditions in which the object exists and changes; (4) the aim, ideals or etalons of this object; and (5) the action of collecting and keeping the tracked information. If we simplify the understanding of *time* and define it as a common and independent scale that tracks the order and the existence of observations, we can define *development* as a set of conditions or the estimations of these conditions of the objects in time.

In this paper, the (positive) bias towards changes or development (e.g., Myrdal, 1974) is withdrawn on purpose to keep the widest possible implementation of the term. For instance, the development of cancer is not a positive change for a human, but still a set of ordered conditions. Another example, is the growth or the decrease—the positive change until the object suffers from its size more than gains (see Sickles & Zelenyuk, 2019). To get a time series, we match the conditions of different objects by the time they were observed and quantified, that is we present the development in a numeric format. *Volatility* occurs when the defined determinants of development (2), (3), (4), or (5) for the object (1) experience changes in time.

2.2 | Volatility, stability, and structural break

In mathematics and statistics, volatility is an estimate of a fluctuation or several fluctuations. The largest body of evidence exists in financial models, where volatility is measured with a set of ordered growth rates, differences, indexes. Repetitive patterns in the development or in the volatility—*cycles or waves*—are the feature of the objects to periodically go through expansion, peak, recession and crisis. The biggest innovation-driven cycles in macro data are observed to have happened since the 1780s, and the periods of these waves last 40−60 years (Kondratiev, 1926; Schumpeter, 1939). Smaller waves are complimentary to the bigger ones and last 15−25, 7−11, and 3−5 years (Juglar, 1862; Kitchin, 1923; Kuznets, 1930). Life cycles of different lengths exist within organizations, institutions, projects, products, within each individual and social group. *Seasonality* is a repetitive pattern in the data of any frequency. *A bubble* is another form of a wave; it has a cumulative character of development in time and finishes with an explosion—a rapid and sharp change in development after a local extreme (see e.g., Sornette et al., 2018). The times and phases of waves and bubbles are important to know, because catching the wrong wave may violate volatility measures and modeling results. *Volatility* for the individual research questions and variable formats (e.g., the deviation from the equilibrium, economic volatility) can be synonymous with mathematical *diversity, variability, dispersion* (e.g., Joël, 2012).

Stability is something opposite to volatility and seems to comprise a level of volatility close to zero. However, in this case, how do we understand a stable heartbeat, stable chemical reaction, stable society, or stable economic growth? This question arises because each researcher understands stability in the context and the values of the research area (see, e.g., ECB, 2012; Egan & Schofield, 2009; Rossi et al., 2013). One can clean trends or seasonality and continue to study changes and outliers. In these cases, stability is a considered trend or seasonality. These considerations need to be explained, and in economics, they are usually explained with life or business cycles or other types of seasonality. However, the considered seasonality may capture other important effects or may not contribute to the data development exactly as much as the researchers assessed. For this reason, it is challenging to study stability if the *etalon* development or condition is not defined. The understanding of the etalon requires a lot of knowledge from the field. In many areas of natural sciences, the expected development or etalon condition of the object is based on previous studies. In social sciences, this is not always the case. The sample average or an average plus the trend is not an estimation of the etalon because the sample can be self-selecting and may cover only a specific part of a general sample.

The estimated trend may change after structural changes. In statistics, a *structural break* is understood as a trend change; in general, *structural break* is a change in the conditions where the object exists. It is important to distinguish the effects of structural breaks and the effects of cofounders by definition and in the models. We refer the reader to Perron (2005) and VanderWeele (2019) for a discussion on *structural breaks* and proper *confounder* selection. A more flexible definition of stability includes an acceptable magnitude of fluctuations, that is, fluctuations do not cross the defined corridors around the agreed level, trend, wave or etalon.

3 | OPTIONS OF MEASURE SELECTION

3.1 | Measures

In this section, we focus on stability and volatility as the dependent variables or as a description of a certain part of a dataset. The measures of fluctuations and deviations can be classified into five groups based on their calculation: relative to previous observations, sample statistics, relative to previous deviations, reordering, benchmarking and signaling (Table 1).

The measures in Group 1 are the growth (deviation) and growth rates (in per cent or in ratios) calculated relative either to previous observations or to the important time of the studied period. The n-difference is a growth of the difference of order n – 1. The use of growth rates or differences allows analysts to handle non-stationary series and to continue studying the series with minor losses in tools, but with the loss of the levels and, sometimes, meaning. These measures are employed for the non-stationary series, even if these series are to be stationary theoretically (e.g., Wang & Tomek, 2007).

Group 2 comprises measures based on statistics (see more in Bedeian & Mossholder, 2000; Edelmann et al., 2017). The variance, coefficients of variation and other descriptive statistics are traditional volatility measures and can be well applicable for data description. They may also become the dependent variables if measured on the unified time windows.

Group 3 of the measures is based on the statistics of previous deviations. These measures collect all the benefits and problems of Groups 1 and 2, including the loss of levels and the dependency on the time window.

TABLE 1 Overview of stability measures

Group	Basis of calculation	Measures	Main characteristic	Strength	Weaknesses
1	Previous observations	Growth, growth rates, n-difference	Short memory	Close to the important moment	- Consider only several defined observations in time- Loss of levels
2	All data (sample statistics)	Standard deviation, variance, distance standard deviation, coefficient of variation, corridor	Sample statistics	Take all observations into account	- Rarely become the dependent variables - Depend on a number of observations
3	Statistics of deviations	Average absolute deviation (or average deviation), average absolute deviation from the median, mean absolute difference (or mean absolute difference), and median absolute deviation	Group 2 applied to Group 1	Analyze all defined deviations	- Loss of levels - Rarely become the dependent variables - Depend on a number of deviations
4	All reordered observations	Gini coefficient, relative mean difference (twice the Gini coefficient), ranges, interquartile range, and quartile coefficient of dispersion, range divided by mid-range	Reordering	Help avoid self-selecting and outliers in the sample	- Drops time information (correlation in time)
5	Benchmark by the Expert's criteria or signaling rule	Dummy variable that benchmarks or signals the stable or unstable observations	Probability	- Allows applying supervised classification methods and superlearner algorithms- Unified	- Drops most information on the magnitude of fluctuations

Group 4 is based on reordering and share calculating (see Lane, 2003, pp. 144–151). Reordering and shares may be applied to panel data with more than one time period in order to track the development of specific changes in the sample. Ranges are also used to produce range-based indexes (e.g., Pinches & Kinney, 1971). For panel data, it is possible to track the Lorenz curve (Gini coefficient) or interquartile range over time. Stabilization will modify into a 45° line for the Lorenz curve or any of the stability measures applied to the interquartile range number.

Group 5 is based on benchmarking. The experts' criteria define the bandwidth of being stable or unstable. The dependent variable is replaced with a dummy variable that benchmarks the stable or unstable observations, whereas the set of explanatory variables defines the conditions when each outcome was observed. This measure allows applying the models that predict the probability of the unstable outcome. In signaling, the expert's criteria relies on data-driven characteristics, the combination of those, propensity scores and the thresholds for them, as well as the stability measure can be selected using principal component analysis. As a result, the benchmarks of the observations can be interpreted as signals and, therefore, be processed using signal processing concepts (see more in Ortega, 2021; Ortega et al., 2018; Vaswani et al., 2018).

Measures of groups 2 and 3 may differ significantly between different parts of the series or samples and depend on a number of observations. This is a measurement problem in statistics, which was also observed for a variety of indexes for diversity and entropy measuring (see Jost, 2006). Therefore, there is no best volatility measure, but the presented weaknesses and strengths may help to choose the most useful ones for the research.

3.2 | Transformations and frequencies

Besides direct volatility measure calculation, several possible transformations of the data may help to study fluctuations (Table 2). The data can be: transformed into 1st difference, combined with another variable, logarithmized, normalized, standardized, indexed on something, de-trended, calculated relative to itself in a different period or to something else. The described measures and transformations are applied on the variables in levels either over time or across the objects. The interpretation of the results then changes with the selected variable type that is why the measures and transformations are not combined with each other and usually are not applied twice. Any of the defined measures can become the dependent variable for the modeling.

Weekly, monthly and yearly observations will in most cases deliver different statistics and stability measures. The lower frequency mutes the fluctuations that occur at higher frequency. This effect will also lead to changes in estimations. A good start for understanding this effect is to accept that the development is not interrupted in time, but the measurements of development in most cases are the average or a sum for changes that occur at higher frequency. Moreover, the measures do not change with the same rule as the one used to change the scales.

One can assess the models at different frequencies, but in most cases, the selected frequency follows the research question or institutional setting. In the environment of financial markets, people may predict the price, for example, for the next 30 min as they usually use tick-by-tick market data. Other studies may average yearly observations for several years (e.g., Burnside & Dollar, 1997) and estimate trends and waves. Data of higher frequency is required if changes occur within the period or aggregation level of one observation. This is particularly the case when analyzing season-sensitive policies (e.g., Loginova et al., 2021). Usually, the higher the selected frequency, the narrower are the confidence intervals for estimations. Besides obtaining more observations,

TABLE 2 The overview of the basic variable transformations (formats)

Transformation	Measures of deviations	Helps to gain	Helps to lose	Keeps
Differences	Measures of group 1	Stationarity, Comparability[b]	SeasonalityTrends Structural breaksLevels	Order over time
Normalization	A variable measured in per cent of the same variable's maximum	Comparability[b]	Levels	Seasonality TrendsStructural breaks
Standardization	A variable measured in per cent of defined level	Comparability[b]	Levels	Seasonality TrendsStructural breaks
Indexation	For the corresponding time and object, the variable is measured relative to the other variable (in percent or ratios)	Normality StationarityInvestigation of two variables	Co-movement Seasonality TrendsStructural breaks	
Combination	For the corresponding time and object, a variable is summed with other variable of the same units	Normality Stationarity Investigation of a sum of two variables	The information about the single variables	Co-movement Seasonality TrendsStructural breaks
Logarithm[a]	A natural logarithm of the defined variable	Discovering elasticitiesChange towards linear model	ScalesLevels	Applicability of the soft for linear modeling
De-trending	A variable is regressed on time and this component is excluded from the variable's values	TS tools Uncertainty[c]	TrendsCo-movement	Applicability of the TS modeling tools
Cleaning seasonality	A variable is regressed on seasonal trend and this component is excluded from the variable's values	TS tools	Seasonality	Levels Uncertainty[c]
Principal component analysis	Measures 1–4 or any transformation using characteristics of observations Measure of homogeneity, structure, relations and robustness of observations	Possibility of big data visualization, Graph Signal tools,Reduction of computational costs	Many layers in machine learning techniques The information about the single variables	Applicability of signal processing concepts The variety of stability understanding

[a]Logarithm is a linear transformation and is applied to the targeted variable plus 1 to keep zeroes in the dataset.
[b]of fluctuations on different levels.
[c]about the excluded components.

this is an important reason why the higher frequency is preferable to the lower one. However, in practice, the available data often defines the frequency selection.

To explain weekly changes with monthly changes, one will have to duplicate, share or extrapolate monthly data for each week of the month and assume supervised variation or its absence at the higher frequency. The higher frequency data help fulfil regular assumptions on the residuals. The more observations one has, the more likely one gets normality of the residuals, whereas other problems of residuals (heteroscedasticity and autocorrelation) in practice are cleaned with the "robust" option of statistical packages. The daily data and the higher frequent data often suffer from "weekend effect" if the data are absent or do not change on Saturday and Sunday.

4 | MODELING OF STABILITY

Table 3 provides a short overview of the most used methods. Researchers measuring the development over time tend to switch from levels to volatilities, especially when a time series analysis is needed. In the 1980s, time series analysis culminated in the (generalized) autoregressive conditional heteroscedasticity ([G]ARCH) models (Bollerslev, 1986; Engle, 1982), which allowed explaining fluctuations with the combination of previous fluctuations, errors of the model and their statistics. Since then, autoregressive integrated moving average (ARIMA), GARCH and vector autoregression (VAR or VEC) families of models have been developed further and have remained amongst the most attractive tools for time series analysis.

Scientists traditionally focus on the background, drivers, the responses and causalities. With a number of strict assumptions, causalities are investigated mostly with causal and experimental methods (see more in Huber, 2019) because these methods have a more precise definition of causality as compared with time-series approaches (Lechner, 2010). Besides the mentioned modeling methods, stability may also be studied with descriptive statistics, ordinary least squares (OLS), panel data models and even Machine learning algorithms (Superlearners and Graph Signal Processing). All the methods require robustness checks, that is, testing the magnitude of the estimates while running the specification at a different time or time window or for a random cut-off or treatment, if any. Data availability always forces the researchers to derive modeling conclusions for *only a studied subsample and time* frame. The studied subsamples in the best case should avoid three types of bias: self-selection bias, undercoverage bias and survivorship bias (see Lane, 2003, pp. 235–237). The milestones of key methodological developments (based on *Number*, *Graph* and other earlier theories) are shown in Figure 2.

DiD and RDD require the causality identification and interference statements, at least Conditional independence assumption (CIA), Stable-unit-treatment-value assumption (SUTVA); Common support assumption (CS) and Exogeneity assumption (EXOG) should be satisfied (see, among others, Lechner, 2019). RDD, VAR, ARIMA, GARCH, OLS and Panel data models require at least 100 observations for the asymptotics. SuperLearners are used on big data. The choice of the method follows the interest. The violation of the assumptions violates the results.

The DID is a powerful causal estimation approach to measure treatment effect, but it requires an approved control group for comparison and must fit the common trend assumption. The main effect is called the "average treatment effect on the treated" and is the measurement of volatility (shift) introduced by the treatment in the treatment group compared to a control group. Stationarity is sometimes required for this method, but the common trend assumption is always necessary and can introduce several problems to the study. One can refer to the literature (Jaeger et al., 2020; Roth, 2019) to see how each observation can significantly change the estimation of the slope of

TABLE 3 The overview of the methods to study stability

Families of models	Variables of interest (DV)	Explanatory variables (EV)	Applied to	The result (the interest)	The main assumptions
Difference-in-differences (DID)	One treated set of observations, similar set of observations not treated, both have the observations before and after treatment	EV influence almost the same for the studied and comparison group. Fixed effects and cofounders may be included.	Time series, panel and individual data with available treaded and control groups before and after treatment	ATET - Causal average treatment effect on the treated (volatility/level shift introduced by the treatment)	The common pre-trend assumption. Causality, identification and interference statement.
Sharp or fuzzy regression discontinuity design (RDD)	One treated series with the observations before and after treatment	EV do not influence during the studied short period. Fixed effects and cofounders may be included.	A narrow, balanced interval around structural break	Causal effect (the shift happened at the defined cut-off point)	The absence of seasonality. A clear cut-off point.Causality, identification.
VAR and its modifications	Several stationary series that interdepend in theory	EV include all possible (best) confounders and their lagged transformations.	Collected data	Series' interdependencies prediction	Stationarity of all series. Stability of the model. OLS assumptions on the residuals.
ARIMA, GARCHAnd their modifications	One stationary series	The influence of other factors is random.	A series that follows speculative behavior or seasonality	Moving seasonality (how current observation may be defined by previous deviations and shocks)	Stationarity of all series.OLS assumptions on the residuals
Panel data models	Panel data with small number of periods and many observations for many objects	All possible (best) confounders.	All observations where a time fixed effect can be defined	Shifts in average levels and the contribution of factors' change to the change in DV	OLS assumptions on the residuals if assessed with linear models

(Continues)

TABLE 3 (Continued)

Families of models	Variables of interest (DV)	Explanatory variables (EV)	Applied to	The result (the interest)	The main assumptions
OLS	Stability measures calculated for one feature of many objects	All possible (best) confounders.	Collected data	The effect of confounders on the DV	All OLS assumptions
Machine learning: Superlearner and Graph Signal Processing	The column in big data defined as stability measure	Big data, usually in panel format.	Collected data	The model or classification that predicts the outcome (DV) with the same set of EV	Usually numeric data without missing values. The models are trained on training dataset and then tested on test dataset

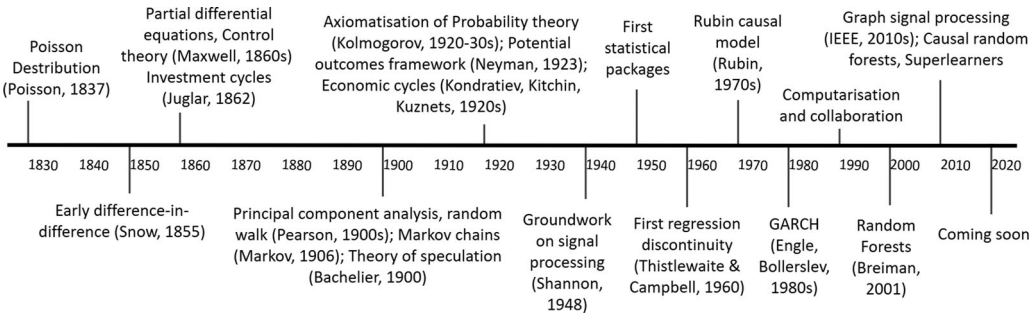

FIGURE 2 *The milestones of the methodological progress over time*. This figure illustrates methodological progress by mapping only a few selected key works on a timeline. For more details, see Bachelier (1900), Bollerslev (1986), Breiman (2001), Engle (1982), IEEE (2010), Juglar (1862), Kitchin (1923), Kolmogorov (1933), Kondratiev (1926), Kuznets (1930), Markov (1906), Maxwell (1868), Neyman (1923), Pearson (1901), Pearson (1905), Poisson (1837), Rubin (1974), Shannon (1948), Snow (1855), Thistlewaite and Campbell (1960)

the trend. The change in a number of observations may help fulfil the common trend assumption; however, if the estimated effects and their significance differ largely with the introduction of a couple of new observations, robustness of the model is failed. A time of structural break is sometimes excluded from data to avoid the influence of the transition period on the estimates (e.g., Loginova et al., 2021). Propensity score matching allows pairing observations from treated and control groups by similar values of characteristics (see the discussion in Caliendo & Kopeinig, 2008).

The RDD measures the shift in the sub-population close to the cut-off (see Imbens & Lemieux, 2008; Imbens & Rubin, 1997; Lee & Lemieux, 2010; Thistlewaite & Campbell, 1960). This method requires (1) avoiding seasonality in the data, (2) explaining the cut-off (consistency) and the model identification (unconfoundedness), (3) ensuring comparable observations within the window of estimation (randomization, positivity), and (4) adoption if applied to time series (Hausman & Rapson, 2018). Similar to the possibility to calibrate the pre-trend with the observations for DID, the RDD results can significantly change with the changed bandwidth. Pooling observations from too far from the cut-off or different years violates the design. A procedure for optimal bandwidth selection exists, but the result of this procedure may depend on the number and the magnitude of observations in the data. The second option for bandwidth selection is an expert's decision that is usually driven by some knowledge about the period of assessment. Lee and Lemieux (2010) advised testing the bandwidths of different lengths instead of changing kernels. The best bandwidth is the one, which is close to the cut-off and provides at least 100 observations for the assessment. Overlapping of the bandwidths for different treatments is not allowed even if the treatments are insignificant. Fuzzy RDD allows the probability that several observations after the cut-off remained unaffected. One can also face a lagged effect of the treatment when the RDD cut-off requires calibration.

ARIMA, GARCH and VAR modeling predicts interdependencies and suffers from non-stationarity of the series, especially regarding time series with policy effects (structural change), when the trends change or the levels shift. Any dummy in time series models does not measure the effect but just measures the shift in the intercept. All-time series must pass the unit root test (reject unit root) if one applies any of the ARIMA components (e.g., Dickey & Fuller, 1979; Zivot & Andrews, 1992). All tests (incl. *p*-values for significance tests) performed on non-stationary series can deliver untrue results because most of the tools, tests and theorems for these models are

developed in an assumption of stationarity. A stationary set of series allows studying interdependencies of fluctuations, price transmission, but not a shock transmission between the series (except impulse response functions) and not causal effects (with a few exceptions, see more in Lechner, 2010). Various tests on stationarity were developed to allow "fishing" the *p*-value, however most time series in levels are not stationary. If the unit root is not rejected, people de-trend or transform the non-stationary series into 1st difference. These models on differences may clean level discontinuities including policy-driven effects. A series integrated of higher order *n* ("I" part in ARIMA) have more chances to pass the tests on stationarity. However, ARIMA often loses explanation after the 2nd integrating. ARIMA explains the current observation with several previous observations ("AR" part) and shocks ("MA" part). This means that the less share of each fluctuation is explained by the set of previous observations ("AR" part), the more will have to be explained with the combination of previous errors ("MA" part) and vice versa. VAR allows AR parts to interrelate between several series. [G]ARCH includes the properties of shocks into the model. This speculative nature of ARIMA models is similar to seasonality but moves in time together with the observations and the errors of the model.

Panel data for several periods often allow studying the development. One can define as dependent variable a stability measure for each period across observations. One can also join the periods into sequential groups and measure stability for each of these groups. Time fixed effects then must be included in the model. It is worth mentioning that both fixed and random effects models should be applied and accompanied with time fixed effects. Recent studies developed tools to assess models with multiple fixed effects (Gaure, 2020) and mixed models containing both random and fixed effects in one equation (Bell & Jones, 2014). Both require slightly more individual observations than classic methods. For OLS and time fixed effects models, the data must pass the tests for residuals. Because current datasets tend to be big and diverse, recent developments in big data analysis and super learning may assist in stability studies if the necessary amount and type of data are available. The Group 5 of stability measures helps apply logit- and probit-models and most of the supervised machine learning techniques, including classification tools, random forests and signal processing concepts.

In the current literature, the mentioned models are refined both within each design (e.g., DID with multiple periods, see Callaway & Sant'Anna, 2020) and by synergies between the designs. Random forests (Breiman, 2001) are increasingly combined with causal methods. For example, random forests have previously been combined with RDD (e.g., Asher et al., 2016) as well as with average, individual and group treatment effects in DID (e.g., causal forests by Athey & Wager, 2019; Lechner, 2019). Time-series methods are less preferable for combining with causal methods as they can displace the effects to the errors of the model. Diverse collections of simple models are shown to be useful for describing complex phenomena (see more in Richerson & Boyd, 1987).

5 | SUGGESTIONS ON MEASURE AND MODEL SELECTION

Standardization or the first difference employed within the groups helps avoid most of the problems of comparability in the data. Each allows comparing different groups of the objects, for example, individuals, products, countries, organizations, their characteristics and estimated coefficients. Big data studies often employ normalization of the data. If the study values the information on the magnitude of the dependent variable same as the impact of drivers, then it is better to avoid using ordered models. If the study's interest is only the magnitudes of the dependent variable than volatility measures of groups 1–4 would be the best option for investigation.

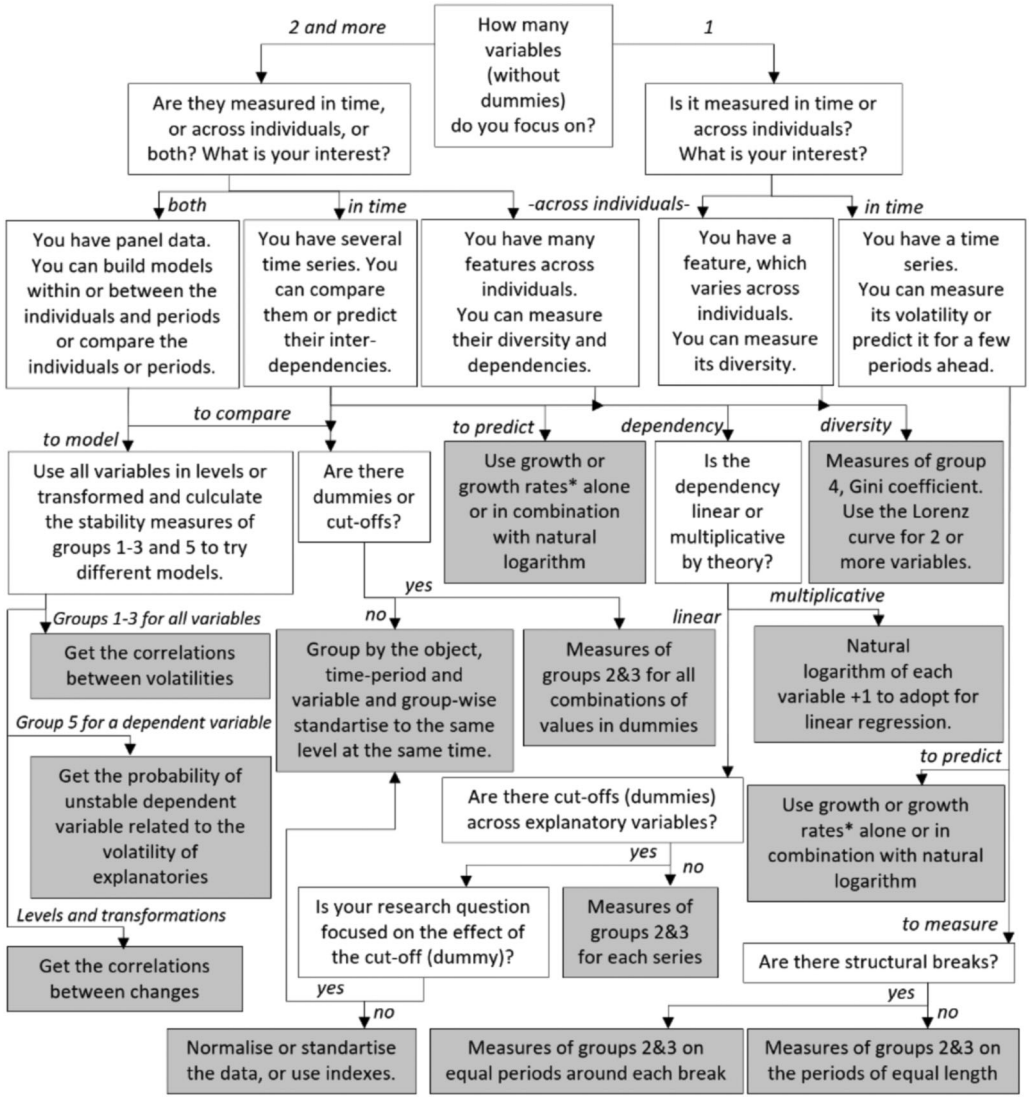

FIGURE 3 The scheme of volatility measure selection. Note. "*" denotes that structural breaks are neglected

Figure 3 illustrates the simplified top-down scheme of volatility measure selection. The measure selection depends mostly on data availability, interest and the presence of structural breaks in the data. The stability of the dependent variable across individuals is better studied with Gini coefficients, while the various measures and transformations can be applied for time series. For tracking the changes in time, Gini coefficients and coefficients of variation are calculated for each period. The Lorenz curve visualizes the deviation of the actual data from 45° line of etalon equality.

To measure stability with a model, one should focus on the particular interest. Figure 4 illustrates the zones of interest $Z_1 \ldots Z_4$, where the stability of the studied series Y may be measured. We define these zones with Z_i. The intercepts of series Y in zone Z_i are $E(Y_{Z_i})$, and the trends are $T(Y_{Z_i})$. The assessments of $E(Y_{Z_i})$, $T(Y_{Z_i})$ and statistics will differ depending on the chosen

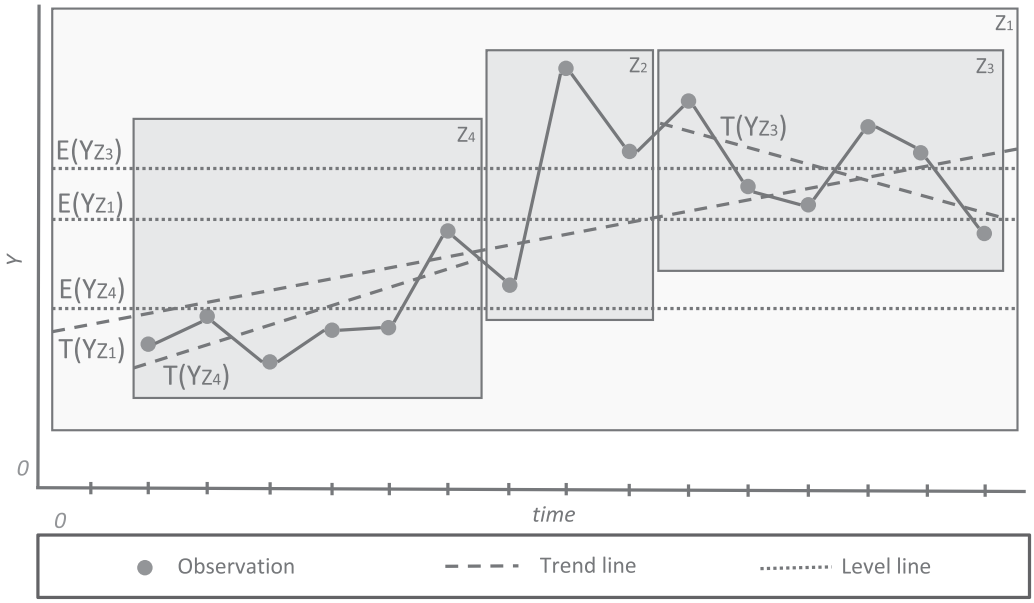

FIGURE 4 An illustration of the zones of stability over time. Note. Each square on the graph collects the observations for the zone of interest denoted in the square's top right apex

focus (zone Z_i) of the study. One can measure stability of all available data (Z_1), to study only the moment of structural break and transition (Z_2) or to consider the structural break and study the differences of zones Z_3 and Z_4.

One can choose any of these zones to study separately or compare them with each other. During comparisons, the equal length of the zones allows employing of powerful measures that are sensitive to the period length, for instance, coefficient of variation. One can benchmark those of the observations fitting some experts' criteria on excessive fluctuations and continue the analysis with the methods of probability estimation for excessive fluctuation.

Traditionally, longer series are preferable because they allow using more explanatory variables, avoiding overfitting, obtaining normality and tracking tendencies. However, structural breaks may violate the models. To consider a possible trend break, researchers sometimes use dummies or interaction terms and then solve the multicollinearity problem. The longer the studied series is, the stronger the ARIMA, GARCH and VAR models will push the shocks, such as the one depicted as Z_2 in Figure 4, to the errors of the model. The value of one observation decreases in the total assessment with the increasing number of observations. If the effect of interest is exactly the one undervalued, it is better to use a causal DID estimation or RDD. The trends of ARIMA, GARCH and VAR models respond to $T(Y_{Z1})$, the results of the DID estimation respond to $T(Y_{Z_3}) - T(Y_{Z_4})$ minus the same change in the control group, whereas the results of the RDD respond to the shift that happened in Z_2 with a defined width and provided high number of observations. In pooled models for Z_1, the constant will be estimated as $E(Y_{Z_1})$; in the models for Z_1 with time fixed effects for zones Z_3 and Z_4 (Z_2 is dropped), the constant will be estimated for each period separately, that is $E(Y_{Z_3})$ and $E(Y_{Z_4})$, where $E(Y_{Z_3}) - E(Y_{Z_4})$ collects treatment, structural break and other uncontrolled effects in Z_2. Another problem is to define a structural break if the time of it is unknown. In practice, Zivot –Andreus test stays a most used tool for this purpose in time series (Zivot & Andrews, 1992).

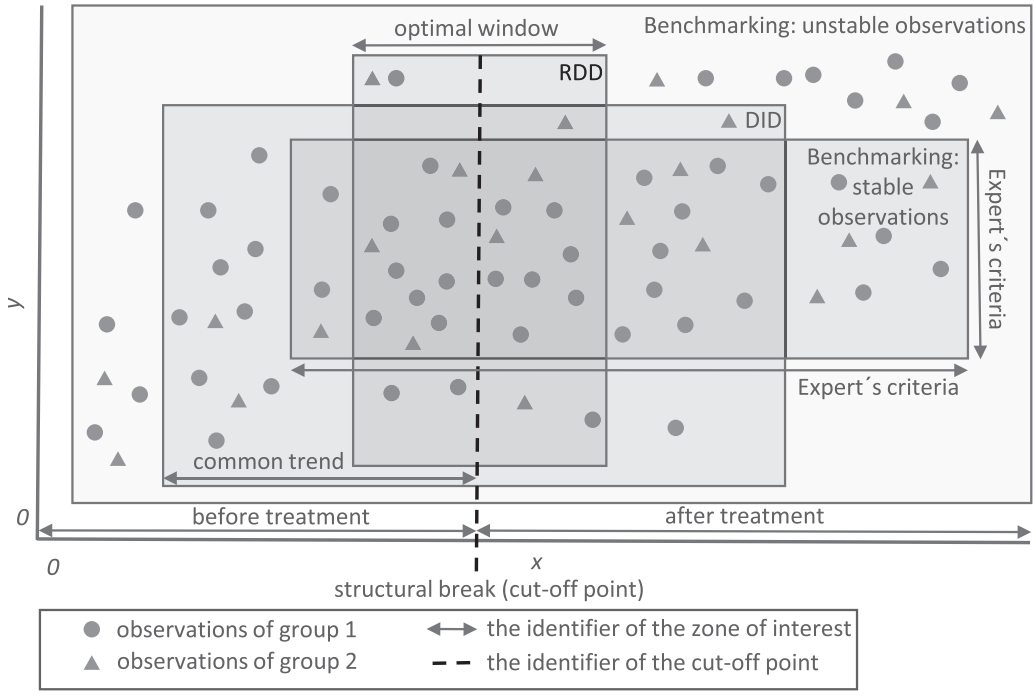

FIGURE 5 An illustration of the zones of stability in groups. Note. Each square on the graph collects the observations for the method denoted in the square's top right apex

Figure 5 demonstrates the zones of interest related to RDD, DID and Benchmarking for panel data with an expert-defined structural break. The techniques of measuring the stability within the groups usually neglect the correlation of the observations in time or assume the absence of this correlation. Therefore, stability across the groups may be broader defined than stability of time series data (e.g., *diversity*).

Classic DID and RDD are specific settings of OLS with treatment dummy, but gain their causal interpretation due to the assumptions. In DID application, the observations of the group 2 will be used as a control group for the observations of the group 1, assuming a common trend in the untreated phase (and other assumptions mentioned in Section 4). If another control group for the groups of observations depicted in Figure 5 exists, then these groups may be distinguished with fixed effects. Benchmarking replaces dependent variable with the dummy to predict with probability estimation methods and SuperLearners. The expert's criteria on stable observation are based on the volatility measures of the group 1.

For RDD, localization of a model is a key for the causal interpretation of estimations. Both DID and RDD allow the inclusion of fixed effects and other confounders, however, the literature recommends using DID rather than RDD whenever possible (Hausman & Rapson, 2018), because the insignificant impact of the other factors in RDD is difficult to prove in practice.

Figure 6 illustrates the simplified top-down scheme of model selection. One may see that the model selection starts with data availability and interest, continues with several data-specific decisions and further concentrates on proper implementation of the methods.

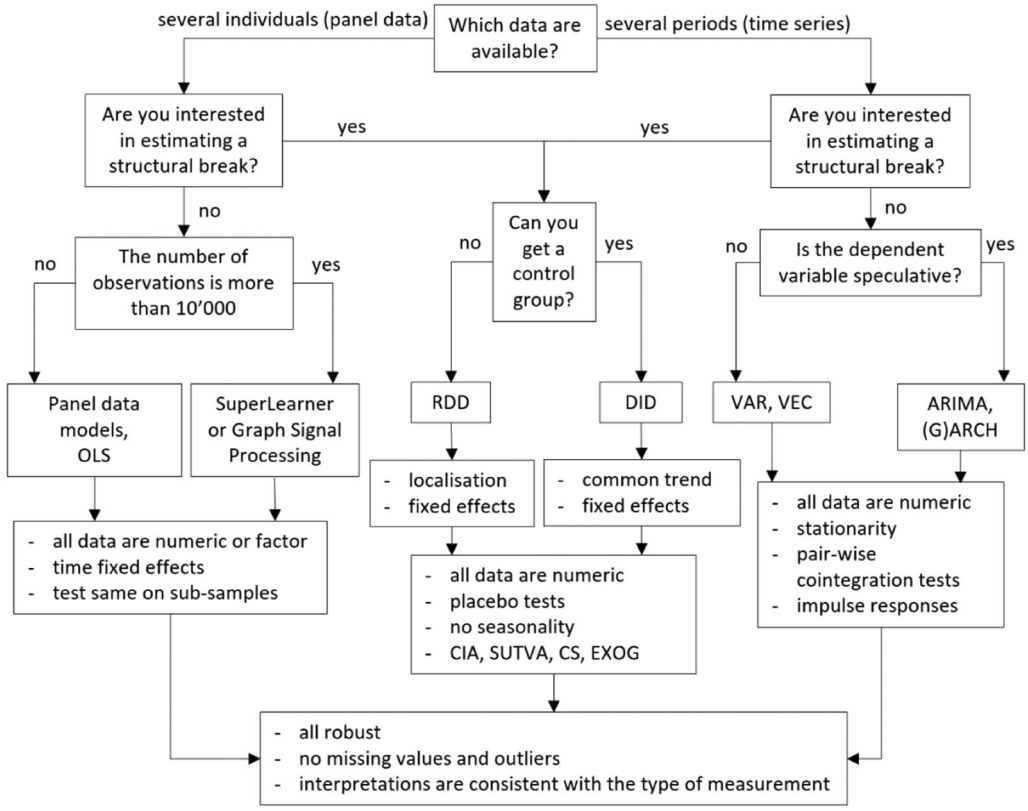

FIGURE 6 The scheme of model selection

6 | CONCLUSION

Stability during and beyond structural breaks is an important research interest, and this paper provides a roadmap and explanation of the basic methods that help to streamline additional stability research across many sectors, countries and research areas. We summarize the basic knowledge and basic methods for measuring volatility and stability, explain their advantages and disadvantages, and suggest ways to investigate volatility and stability under different circumstances. The reader is encouraged to develop, combine and apply modifications to the discussed tools if the design and measurement are appropriate for the research task.

The paper shows that there is no "one size fits all" solution but the suitable measurements and methods strongly depend on the scope of the data and the objective of analysis. As discussed in the paper, stability measurement and modeling can be challenging, especially at the stage of selecting the model and meeting its assumptions. Therefore, we see further challenges and perspectives in the field of modeling, especially in optimizing and improving algorithms, machine learning and facilitating modeling assumptions. Furthermore, the variety of research tasks in which the discussed measurements and models can be applied remains infinite. Therefore, the study of stability remains a promising avenue for further research.

ACKNOWLEDGMENTS

Open access funding provided by Agroscope.

CONFLICTS OF INTEREST

We have no conflicts of interest to disclose.

DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

ORCID

Daria Loginova https://orcid.org/0000-0002-4856-9648

REFERENCES

Asher, S., Nekipelov, D., Novosad, P., & Ryan, S. (2016). *Classification trees for heterogeneous moment-Based models*. NBER Working Paper 22976. http://www.nber.org/papers/w22976

Athey, S., & Wager, S. (2019). *Estimating treatment effects with causal forests: An application*. arXiv:1902.07409.

Athey, S., & Imbens, G. (2019). *Machine learning methods economists should know about*. arXiv:1903.10075.

Bachelier, L. (1900). Théorie de la spéculation. *Annales Scientifiques de l'École Normale Supérieure*, *3*(17), 21–86.

Bedeian, A. G., & Mossholder, K. W. (2000). On the use of the coefficient of variation as a measure of diversity. *Organizational Research Methods*, *3*, 285–297. https://doi.org/10.1177/109442810033005

Bell, A., & Jones, K. (2014). Explaining fixed effects: Random effects modelling of time-series cross-sectional and panel data. *Political Science Research and Methods*, *3*, 133–153. https://doi.org/10.1017/psrm.2014.7

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *31*, 307–327. https://doi.org/10.1016/0304-4076(86)90063-1

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A:1010933404324

Burnside, C., & Dollar, D. (1997). Aid, policies, and growth. Policy research working paper 1777. The World Bank. https://documents1.worldbank.org/curated/en/698901468739531893/pdf/multi-page.pdf

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, *22*, 31–72. https://doi.org/10.1111/j.1467-6419.2007.00527.x

Callaway, B., & Sant'anna, P. H. C. (2020). Difference-in-differences with multiple time periods. *Journal of Econometrics*, *225*, 200–230. https://doi.org/10.1016/j.jeconom.2020.12.001

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, *74*, 427–431. https://doi.org/10.2307/2286348

ECB (2012). *Financial stability: Measurement and policy. Frankfurt am Main*. https://www.ecb.europa.eu/press/key/date/2012/html/sp120614_1.en.html

Edelmann, D., Richards, D., & Vogel, D. (2017). *The distance standard deviation*. https://arxiv.org/abs/1705.05777

Egan, W., & Schofield, T. (2009). Basic principles of stability. *Biologicals*, *37*, 379–386. https://doi.org/10.1016/j.biologicals.2009.08.012

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, *50*, 987–1007. https://doi.org/10.2307/1912773

Gaure, S. (2020). *lfe R-package*. https://www.rdocumentation.org/packages/lfe/versions/2.8-5.1/topics/felm

Greene, W. (2011). *Econometric analysis* (7th edn.). Prentice Hall.

Hausman, C., & Rapson, D. (2018). *Regression discontinuity in time: Considerations for empirical applications*. NBER Working Paper Series No 23602. https://www.nber.org/system/files/working_papers/w23602/w23602.pdf

Huber, M. (2019). *An introduction to flexible methods for policy evaluation*. Working Papers SES. University of Freiburg. https://arxiv.org/pdf/1910.00641.pdf

IEEE (2010). Electronic resources in the field of graph signal processing. https://web.media.mit.edu/~xdong/resource.html

Imbens, G. W., & Rubin, D. B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *Annals of Statistics*, *25*, 305–327. https://doi.org/10.1214/aos/1034276631

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, *142*, 615–635. https://doi.org/10.1016/j.jeconom.2007.05.001

Jaeger, D., Joyce, T., & Kaestner, R. (2020). A cautionary tale of evaluating identifying assumptions: Did reality TV really cause a decline in teenage childbearing? *Journal of Business & Economic Statistics*, *38*, 317–326. https://doi.org/10.1080/07350015.2018.1497510

Johnson, D. (2008). *Marketing studies and market considerations. Fundamentals of Land Development*. John Wiley & Sons. https://doi.org/10.1002/9780470260043

Jost, L. (2006). Entropy and diversity. *Oikos*, *113*, 363–375. https://doi.org/10.1111/j.2006.0030-1299.14714.x

Joël, C. (2012). *Measuring macroeconomic volatility. Applications to export revenue data, 1970–2005*. Working paper I14. https://ferdi.fr/dl/df-ffDuTsM2SZQq6ftuTVVBGV5C/ferdi-i14-measuring-macroeconomic-volatility.pdf

Juglar, C. (1862). *Des crises commerciales et leur retour periodique en France, en Angleterre, et aux Etats-Unis*. Guillaumin. https://doi.org/10.4000/books.enseditions.1382

Kitchin, J. (1923). Cycles and trends in economic factors. *The Review of Economics and Statistics*, *5*, 10–16. https://doi.org/10.2307/1927031

Kolmogorov, A. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Technische Universität, Mathematische Fachbibliothek.

Kondratiev, N. (1926). Die langen Wellen der Konjunktur. *Archiv für Sozialwissenschaft und Sozialpolitik*, *56*, 573–609.

Kuznets, S. (1930). *Secular movements in production and prices: Their nature and their bearing upon cyclical fluctuations*. Houghton Mifflin Company.

Lantz, B. (2019). *Machine learning with R* (3rd edn.). Packt Publishing Ltd.

Lane, D. (2003). *Introduction to statistics (online ed.)*. https://onlinestatbook.com/Online_Statistics_Education.pdf

Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, *48*, 281–355. https://doi.org/10.1257/jel.48.2.281

Lechner, M. (2010). The relation of different concepts of causality used in time series and microeconometrics. *Econometric Reviews*, *30*, 109–127. https://doi.org/10.1080/07474938.2011.520571

Lechner, M. (2019). *Modified causal forests for estimating heterogeneous causal effects*. arXiv: 1812.09487.

Loginova, D., Portmann, M., & Huber, M. (2021). Assessing the effects of seasonal tariff-rate quotas on vegetable prices in Switzerland. *Journal of Agricultural Economics*, *72*, 607–627. https://doi.org/10.1111/1477-9552.12424

Maggino, F., & Facioni, C. (2017). Measuring stability and change: Methodological issues in quality of life studies. *Social Indicators Research*, *130*, 161–187. https://doi.org/10.1007/s11205-015-1129-9

Markov, A. (1906). Rasprostranenie zakona bol'shih chisel na velichiny, zavisyaschie drug ot druga. *Izvestiya Fiziko-matematicheskogo obschestva pri Kazanskom universitete*, *2*(15), : 135–156.

Maxwell, J. C. (1868). On Governors. *Proceedings of the Royal Society of London*, *16*, 270–283. https://doi.org/10.1098/rspl.1867.0055 JSTOR 112510

Myrdal, G. (1974). What is development? *Journal of Economic Issues*, *8*, 729–736. http://www.jstor.org/stable/4224356

Neyman, J. (1923). *Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes*. Master's Thesis. Excerpts reprinted in English, *Statistical Science*, *5*, 463–472. (D. M. Dabrowska, and T. P. Speed, Translators).

Ortega, A., Frossard, P., Kovacevic, J., Moura, J. M. F., & Vandergheynst, P. (2018). Graph signal processing: Overview, challenges, and applications. *Proceedings of the IEEE*, *106*(5), 808–828.

Ortega, A. (2021). *Introduction to graph signal processing*. Cambridge University Press.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, *2*(11), 559–572. https://doi.org/10.1080/14786440109462720

Pearson, K. (1905). The problem of the random walk. *Nature*, *72*(1865), 294. https://doi.org/10.1038/072294b0 S2CID 4010776

Perron, P. (2005). *Dealing with structural breaks. Palgrave handbook of econometrics (Vol. 1)*. Boston University. http://www.zafzaf.it/clima/cm28/perron.pdf

Pinches, G. E., & Kinney, W. R. (1971). The measurement of the volatility of common stock prices. *The Journal of Finance*, *26*, 119–125. https://doi.org/10.2307/2325745

Poisson, S. (1837). *Probabilité des jugements en matière criminelle et en matière civile, précédées des règles générales du calcul des probabilités*. Bachelier.

Rabie, M. (2016). *Meaning of development. In A theory of sustainable sociocultural and economic development*., Palgrave Macmillan. https://doi.org/10.1007/978-1-137-57952-2_2

Richerson, P., & Boyd, R. (1987). Simple models of complex phenomena: The case of cultural evolution. In J. Dupre (ed.) *The Latest on the best: Essays on evolution and optimality*. MIT Press. pp. 27–52.

Roth, J. (2019). *Pre-test with caution: Event-study estimates after testing for parallel trends*. https://scholar.harvard.edu/files/jroth/files/roth_pretrends_20190730.pdf

Rossi, B., Elliott, G., & Timmermann, A. (2013). *Advances in forecasting under instability*. Handbook of Economic Forecasting, Elsevier Publications.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. https://doi.org/10.1037/h0037350

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*, 379–423, 623–656.

Schumpeter, J. (1939). *Business cycles*. McGrau-Hill.

Sickles, R., & Zelenyuk, V. (2019). *Measurement of productivity and efficiency: Theory and practice*, Cambridge University Press. https://doi.org/10.1017/9781139565981

Snow, J. (1855). *On the mode of communication of cholera*. London: John Churchill.

Sornette, D., Cauwels, P., & Smilyanov, G. (2018). Can we use volatility to diagnose financial bubbles? Lessons from 40 historical bubbles. *Quantitative Finance and Economics*, *2*, 486–590. https://doi.org/10.3934/QFE.2018.1.1

Thistlethwaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex-post facto experiment. *Journal of Educational Psychology*, *5*, 309–317. https://doi.org/10.1037/h0044319

Vanderweele, T. J. (2019). Principles of confounder selection. *European Journal of Epidemiology*, *34*, 211–219. https://doi.org/10.1007/s10654-019-00494-6

Vaswani, N., Bouwmans, T., Javed, S., & Narayanamurthy, P. (2018). Robust subspace learning: Robust PCA, robust subspace tracking and robust subspace recovery. *IEEE Signal Processing Magazine*, *35*(4), 32–55.

Wang, D., & Tomek, W. G. (2007). Commodity prices and unit root tests. *American Journal of Agricultural Economics*, *89*, 873–889. http://www.jstor.org/stable/4492867

Wooldridge, J. (2013). *Introductory econometrics: A modern approach* (5th edn.). Cengage Learning.

Zivot, E., & Andrews, D. W. K. (1992). Further evidence on the great crash, the oil-price shock, and the unit-root hypothesis. *Journal of Business and Economic Statistics*, *10*, 251–270. https://doi.org/10.2307/1391541

How to cite this article: Loginova, D., & Mann, S. (2023). Measuring stability and structural breaks: Applications in social sciences. *Journal of Economic Surveys*, *37,* 302–320. https://doi.org/10.1111/joes.12505