Describing Location Shifts with One Class Support Vector Machines

igescu, iulia

National Bank of Romania

13 February 2020

Describing Location Shifts with One-Class Support Vector Machines

Iulia Igescu*

February 13, 2020

Abstract

The evolution of variables during location shifts (structural breaks) is of high interest to policy makers. I propose a novel approach to describe location shifts. I use two business surveys in the industry sector (faster soft indicators) to target the industrial production index (a slower hard indicator). Then I use One-Class Support Vector Machines on combinations of these two variables to identify if new observations act as 'novelties' for the target variable, as observations coming from a different distribution. In that case, one would expect the onset/end of a location shift. Moreover, that gives insights into what role animal spirit, as manifested in survey data, plays in equilibrium formation (location shifts).

*With support from National Bank of Romania. The usual disclaimer applies - the views expressed in this paper do not involve National Bank of Romania. Contact: iulia.igescu@bnro.ro

1 Motivation

Economic downturns are a main source of macroeconomic instability. They are especially hard to forecast. According to Clemens and Hendry (2001) that is also due to deterministic shifts in variables. There are frequent small shifts, also called trends (in econometrics this is an I(1) process or a unit root). There are also large shifts occurring mostly at irregular times, shifts in equilibrium levels. They are called location shifts (this is an I(2) process). These two non-stationarity forms are related (see Rappoport and Reichlin, 1989), empirically hard to identify, and have similar solutions: unit roots and mean shifts can be removed by differencing. Cointegration can also remove trends by taking linear combinations of trending variables. Hendry (1995) introduces the concept of co-breaking to deal with location shifts. Just like in the case of cointegration, regime shifts cancel across linear combination of variables such that transformed variables do not depend on breaks. Co-breaking in turn would improve forecasts by getting rid of non-stationarity. Co-breaking could also happen coincidentally because of coincidentally offsetting effects. Luck cannot persist all the time and poor forecasts would eventually resurface when regimes shift again.

While forecasting remains central to policy decisions, forecast performance alone is a misleading criterion for model choice, unless the sole objective is short-term forecasting. Equally important for policy analysis during downturns is to explain how these breaks evolve and when they reach a turning point. Castle, Clements and Hendry (2016) propose to quantify the evolution of variables as the break unfolds, when there is partial information of the changes that are taking place. According to these authors, one should exploit the fact that in practice a fall in the equilibrium mean will alter the growth rate of the variables, as their relationship is not variance free. These changes might take time to complete in dynamical systems. Therefore location shifts might not produce an observable step, but a smoother response that takes time to show up. As impacts of breaks can be delayed, inter-temporal co-breaking must be also considered. That makes it even harder to identify a location shift.

In this paper I select soft indicators that could explain changes in a target indicator and describe their evolution as the break unfolds. At this stage, I do not concentrate on forecasting the size of the shift of the target variable. The goal is to better understand the mechanism of regime shifts for policy purposes. Ideally one would like an early identification of a location shift. Each period, I make a classification assessment if the shift shall continue or not.

Unlike in co-breaking, where a linear combination of variables can in fact eliminate the break to improve forecasts, I use Support Vector Machines (SVM). SVM finds a combination of samples from the current distribution to build a hyper-plane (or a set of hyper-planes in a high dimensional space) to maximize the margin between the old observations and the newest observation in order to separate it from the old observations, if it is 'unusual'. Intuitively, the higher the margin, the better the separation achieved by the hyper-plane.

The SVM algorithm is a nonlinear generalization of the Generalized Portrait algorithm developed in Russia in the sixties. It is grounded on the framework of statistical learning theory that characterizes properties of learning machines which enable them to generalize to unseen data. Smola and Schoelkopf (2004) offer a good introduction into this topic.

The variable I take into account is *industrial production index*. Industry has been going through a prolonged process of change since 2008, materialized in many structural breaks. It has been a source of growth weakness in Europe and the United States. Moreover, the institutional structure of the production process in industry has changed. It has become integrated across economies through super-star firms (value-added chains). Changes in the industry sector in one country feed into economies along the value-added chain of that sector. That in turn induces changes in the country of origin with a lag, making variables quantifying industrial production prone to co-breaking.

I use soft indicators to describe the evolution of the industry production index as a break unfolds. There is a wealth of business survey indicators collected each month. The advantage is that they are faster than targeted hard indicators. Moreover, they give information along many dimensions of the industry sector, such as recent trends, expectations about future trends, employment, prices, export orders, and domestic orders. The disadvantage is that they are not really time series per-se, they are just a sum of opinions, a manifestation of 'animal spirit'. It is well known from Keynes the importance of animal spirit in economy. Farmer and Woodford (1997) prove theoretically that animal spirit is in fact a fundamental, just like capital and labor. We are still in early stages in quantifying information to understand the evolution of animal spirit during structural changes. Therefore this is an added benefit of the current paper, as it also gives some insights into the evolution of animal spirit in the equilibrium formation process.

At first sight, one would expect these surveys to act as leading indicators. Clements and Hendry (op.cit) point out that leading indicators are often not causally related to the variables they lead and if they are subject to breaks, their relationship is unlikely to co-break. It has become common practice for financial institutions to perceive them as leading indicators, treating a weakening in survey data often as a signal of bad times ahead and making investment decisions based on their latest observations. Scatterplots show that in fact survey indicators have different regimes, similar to those of hard indicators they target. One should first identify the current regime of a soft indicator. The relationship between the hard indicator (the target) and the soft indicator (the observable) is therefore not a simple cointegration. Soft indicators selected in this paper are in fact cointegrated with the hard indicator for a period of time, then they have breaks induced by the target indicator, and then often several breaks of their own, at which point their relationship with the target indicator and the relationship among soft indicators becomes circular. It is similar to the behavior of inter-temporal co-breaking, as described by Clemens and Hendry (op.cit.). That in turn would qualify soft indicators as a valid source of information in understanding the evolution of a location shift in a hard indicator. That also points out that for policy purposes, decisions should not be made based on soft indicators latest observation.

I consider soft indicators as a materialization of target indicator changes along various economic dimensions. At this stage I take only on two dimensions, therefore two soft indicators. The first indicator is *production trend observed in recent months* (from now on I will refer to it as recent trend) as a proxy for internal changes. The second is *assessment of export order-book levels* (from now on I will refer to it as exports), as a proxy for external changes coming along the value-added chain. I will use One-Class Support Vector Machines (SVM) introduced by Schoelkopf et al. (2000). In two dimensions, it would be easier to grasp the importance of each dimension every period. Data come monthly from the European Commission. I have chosen data on Romanian economy, as a case where external changes coming from the 'center' of the value-added chain induce domestic restructuring, captured by survey indicators across time. These internal changes would then reflect in the structure of exports. This is a typical case of how modern industry operates in European Union today. It is also a case of intertemporal co-breaking.

2 The Basic Idea

Suppose we have n observations from the same distribution described by p-features. In this paper, take p=2, as there are two soft indicators. Add now one more observation to the data set. If there are no outliers in the old distribution, one would try to detect if the newest observation measured across these two dimenstions is in fact unusual, i.e. it comes from a different distribution. This is called novelty. Novelties can even form a cluster if they are in a low density region of the old data. Old data are used for training.

Mathematically speaking, a density exists if the underlying probability measure possesses an absolutely continuous distribution function. One-Class SVM has been introduced by Schoelkopf et al. (2000). They propose an algorithm that computes a binary function to capture regions in the input space where the probability density lives (to capture its support), i.e. a function such that most of the data will live in the region where the function is nonzero. Moreover, it is applicable also in cases where the density of the data distribution is not even well-defined, e.g. if there are singular components. In situations where the goal is to detect novelties, it is not always necessary to estimate a full density model of the data.

Unlike in the classical case of regression which looks at how many training points fall into the region of interest, this algorithm does the opposite. The algorithm starts with the training points that are supposed to fall into the region, and then estimates a region with the desired property. Often, there will be many such regions. Therefore one has to impose for the region to be small. On a technical note, the measure of smallness depends on the kernel used.

According to Schoelkopf et. al. (op.cit.) to define a frontier for the region one requires a kernel and a scalar parameter, ν. The parameter corresponds to the probability of finding a new, yet regular, observation outside the frontier (this is the error parameter). There is no exact formula or algorithm to set the band-width parameter ν.

The kernel allows for a much larger class of functions by nonlinearly mapping into a high-dimensional feature space. The authors use RBF (Radial Basis Function) kernel. RBF has two parameters: C and γ.

The parameter C trades off accuracy against simplicity of the decision surface. A high C aims at classifying all training examples correctly, making the decision surface look rough. For larger values of C, a smaller margin will be accepted if the decision function is better at classifying all training points correctly. A lower C will encourage a larger margin, therefore a simpler decision function, at the cost of training accuracy. In other words C behaves as a regularization parameter in the SVM.

The γ parameter is the inverse of the radius of influence of samples selected by the model as support vectors. The lower γ, the more influence has a training sample that is far away. When γ is very small, the model cannot capture the complexity or "shape" of the data. The region of influence of any selected support vector would include the whole training set. The resulting model will behave similarly to a linear model with a set of hyperplanes that separate the centers of high density of any pair of two classes.

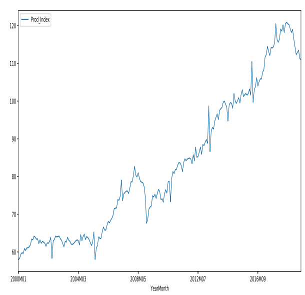
It is therefore clear that calibration of C and γ plays an important role. At this stage, I am taking the calibration recommended by Schoelkopf et al (op.cit.).

3 Assessing Survey Data

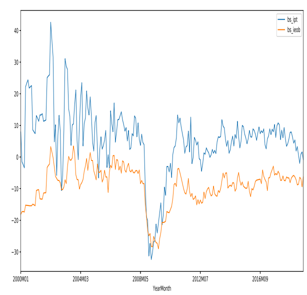
I use two soft indicators, *Production trend observed in recent months* as a measure of internal factors and *Assessment of export order-book levels* as a measure of external factors affecting production. Data are from Eurostat, monthly. The target indicator is *Industry production index*, also monthly. One peculiarity of the survey data is that in periods of change, as it is the case of structural breaks, they tend to be very volatile. On one hand, Romanian industry sector has been going through a complex process of change in the institutional structure of its production process; from a command economy to an integral part of modern super star firms. On the other hand, it moved from labor intensive, to capital intensive, and recently to an information intensive output in only 30 years. That makes Romanian industry sector ideal for studying location shifts. One would also expect to better understand the evolution of animal spirit during such changes. If animal spirit was to precede change, information could be harnessed to understand location shifts. It could therefore become a valid tool for policy analysis.

I would expect to see periods of structural breaks, associated here with supply side changes, accompanied by higher uncertainty, therefore by increased volatility in the soft indicators. A first look at the data for 2000-2019 indicates clearly different regimes in the dynamics of both soft and hard indicators. These regimes seem to have common sources of dynamical changes, making them a valid choice for a detailed look at the evolution of location shifts.

Soft indicators experienced in early 2000s rather wild volatility, in contradiction to an almost flat production index. Change in the industry sector was deep and fundamental, it was institutional. It was an economy where production process moved to new ways of organizing work, while preparing itself for European Union membership. That in turn explains why survey indicators were so violently volatile - it was a very hard period for Romanian economy, of intense institutional experimenting and high uncertainty. As a result, Romania relied on labor intensive industries as main source of exports, a typical evolution for a country with abundant and cheap labor, lacking capital.



[hard]



[soft]

Figure 1: Industrial production and Soft Indicators 2000-2019

After 2004, with ascension treaty to European Union signed, soft indicators decreased in volatility almost over night. Industrial production had its first (and impressive) increase. It seems survey indicators are over-sensitive to institutional changes. They are a measure of trust in the future of an economy, as Farmer and Woodford (op.cit.) have pointed out. They have to do more with the future than the present.

During Great Recession, both soft and hard indicators experienced the same massive and abrupt drop seen all over the world during 2008-2009. There was no warning coming from the soft indicators. There was no wild up and down change, as it had happened during the pre-ascension period. They just kept going down, month after month. That in turn shows that soft indicators could in fact generate 'self-fulfilling' prophecies, as Farmer (1999) proves. A Rare Event feeds into believes, which in turn feed into even lower output levels, moving output to a bad equilibrium. Massive and timely interventions coming from 'center' economies restored trust and by 2010 both hard and soft indicators had fully recovered even at the end of the value-chain, Romania included. This episode is an indicator of how dependent these economies are of each other today. It speaks also about the importance of trust along the value-added chain.

Since then, industrial production index in Romania began its impressive upward ascension till 2018. Soft indicators in turn had moved into even lower volatility regimes. Since 2018, the industry production has been experiencing a protracted decrease, similar only to that of 2008. This decrease has been this time smoother. Soft indicators did not have the wild drop of 2008. Their movements were in fact hard to detect, as they slowly edged towards lower levels of confidence. One explanation for this peculiar evolution is that the economy had previously prepared for this big change through some smaller structural breaks. The indicator *Production trend observed in the last three months* had a similar evolution to that of industry index, moving slower to lower levels. It became negative in August 2019, for the first time since 2013. *Assessment of export order-book levels* on the other hand had been weaker since 2017. It had small ups and downs. Since 2018 it slightly worsened, showing the limits of using only one soft indicator in detecting location shifts. One would therefore have to find more than one relevant indicator to classify the evolution of the industrial production index during an equilibrium change.

These small changes make it hard to interpret the role of soft indicators during structural breaks. Formal structural break tests on all indicators could help us shed some light.

3.1 Structural Breaks since 2010

I consider only breaks after 2010, after European Union ascension and after Great Recession. Formal tests indicate major breaks in 2012-2013, 2015-2016 and 2018-2019. Graphically, peculiar one-period spike-like movements in industry production index seem to have signaled 2012-2013 and 2015-2016 breaks beforehand. That is easy to spot in the graph above. However, the break of 2018 does not have such a spike. Instead industry production index moves downwards through small ups and downs. That is very hard to detect in a timely manner. One explanation could be that these breaks were in fact really different in their dynamical properties, making the exercise of describing the evolution of indicators during breaks even more important.

3.2 'Great Restructuring' Industry Breaks of 2018-2019

TABLE 1
Breaks in Industry Production Index
(2018-2019)

	Break years	
	2018M3	2018M10
intercept	positive	-
trend	negative	negative
type of break	fast	fast

Let us look at soft indicator *assessment of export order level* breaks. This one had in fact a similar negative break in intercept in March 2018, a contemporaneous break, which is what one would expect. The hard indicator continued

[1] Data source for all regressions: European Commission

with another negative shift in October 2010. However, Exports soft indicator reacted with a second shift only in May 2019.

TABLE 2
Breaks in Assessment of Export Order Level
(2017-2019)

	Break years		
	2017M5	2018M3	2019M5
intercept	positive	negative	negative
trend	-	-	-
type of break	slow	fast	slow

On the other hand, breaks in recent trends were detected only in April 2019. A second break, further negative, was in August 2019. By that time, industrial production index had began to visibly show negative effects on the overall GDP growth.

TABLE 3
Breaks in Production trend observed in recent months
(2018-2019)

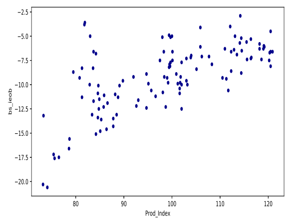
	Break years	
	2019M4	2019M7
intercept	negative	negative
trend	-	-
type of break	fast	slow

The 'center' of the value-added chain induced this process of change, as reflected in exports. European industry sector major changes, mainly in Germany and France, confirm that. It also shows the high degree of integration of Romanian industry to that of the Euro Area. Later on, this process of change had spillover effects on other sectors of Romanian economy. One internal factor aiding to this process of change has been the steady wage increase in Romania since 2015-2016 breaks, at rates of 10-15% year-on-year (yoy). That in turn forced super-star firms to re-organize the production process across the value-added chain, and made investment in more productive technologies profitable. As a result, productivity increased along the value-added chain, and wages increased further in Romania.

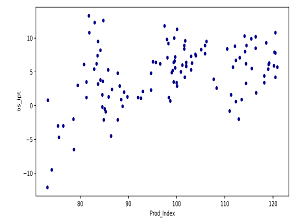
There seems to be a rather intertemporal relationship between variables during this process of change. I will turn to scatterplots, looking for some additional information.

3.3 Scatterplots

Scatterplots indicate a peculiar relationship between industrial production index (depicted on horizontal axis below) and soft indicators. During structural breaks scatterplots seem 'squeezed' during 2012-2013 and 2015-2016 breaks.



[exports]



[recent trend]

Figure 2: Scatterplots of Industrial Production and Soft Indicators 2010-2019

3.3.1 Subsample 2014-2019

Let us have a closer look at the data after 2014. We know that this period includes 2015-2016 breaks.

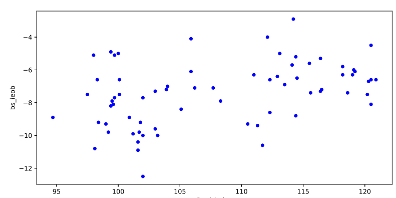


Figure 3: Industrial Production and Exports 2014-2019

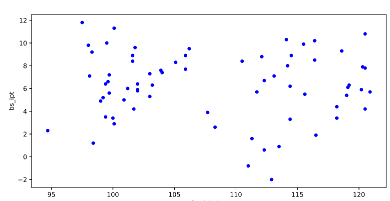


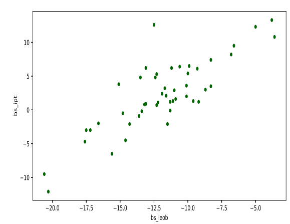
Figure 4: Industrial Production and Recent Trend 2014-2019

Scatterplots of industrial production versus soft indicators indicate an interesting pattern: data are 'squeezed', there is a 'distance' between clouds where the break period is. That is even more obvious in the case of recent trend indicator. Industrial production index is on the horizontal axis in the graphs below.

3.3.2 Soft Indicators over Time

Noteworthy is that the relationship between soft indicators has changed over time. Before 2013 there was a clear sign of contemporaneous cointegration between soft variables, with exports driving change. The break of 2012-2013 is what I call the period of 'old-age rheumatic pain', a continuation of a change process in old ways of production from the early 2000s. Great Recession helped to speed up change. That could explain why industrial production recovered so fast. It increased by 40% in 2012-2018. In general, the evolution of industrial production in Romania in this period is astonishing. After 2013, the relationship between these two soft indicators has again changed. Scatterplots show again the same unusual circular pattern. External and internal factors were feeding into each other inter-temporally, as Romanian industry became more integrated with that of the Euro Area (export indicator is on horizontal axis.).

[before 2013]



[after 2013]

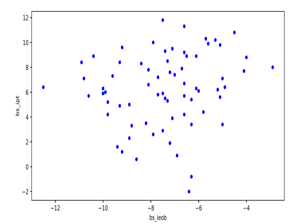


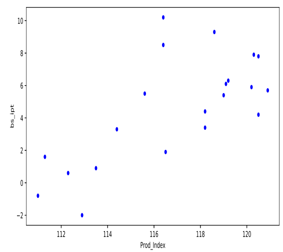
Figure 5: Scatterplots of Soft Indicators Before and After 2013

3.3.3 Scatterplots for 2018-2019 Breaks

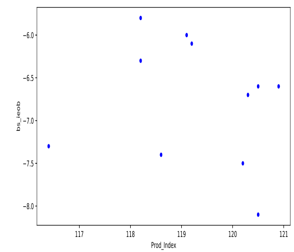
These circular relationships are a sign of intertemporal co-breaking. If one narrows down the sample from the last break of 2016, one could better see the same shape of 'circles' in the middle, corresponding to 2018-2019 breaks. In the graphs below Industrial Production is on the horizontal axis. These 'clouds', this 'distance' between different regimes is then exploited by SVM to build a hyper-plane to maximize the margin between the old data and the newest observation, indicating a possible regime shift.

4 Detecting Novelties with One-Class SVM

To describe the evolution of the industrial production index since 2018, I use One-Class Support Vector Machines (SVM) to clasify data as novelties each period. At this stage, I am taking the calibration recommended by Schoelkopf



[recent trend]



[exports]

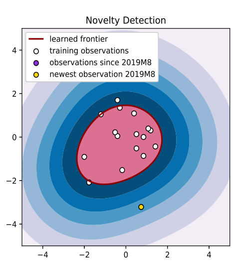
Figure 6: Scatterplots of Industrial Production and Soft Indicators 2018-2019

et al (op.cit.). In this case ν is 0.1, γ is 0.1 and C is 0.5. It is important to decide the sample of training values versus that of testing values.

4.1 Training sample April 2018 to July 2019

As soft and hard indicators break at different times, I will first include in the training sample values from April 2018 (after the break in the hard indicator and in the export soft indicator) up to July 2019, the last break in the recent trend soft indicator. I will take as testing sample values from August 2019 on. As all data points are in fact outside the old regime, they should all be classified as novelites. The system learns the frontier and then sorts out observations. I re-scale the sample to make data from the two different samples comparable. I use the same rescaling parameters for the testing sample. The training sample has 16 values. Only two observations were classified as errors, therefore are in the range of 1%. They are close to the learned frontier in the dark blue area (see graphs below). Training observations are in white. The learned frontier is in red. External factors (*Assessment of Export Order Level*) are on the horizontal axis.

[August]



[September]

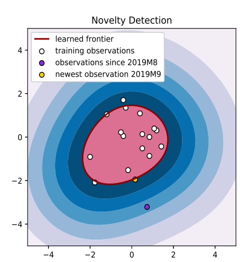
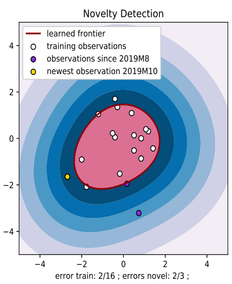
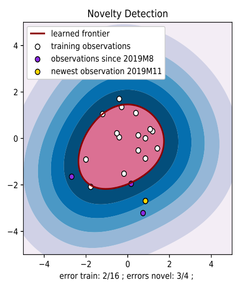


Figure 7: Novelty August; Not-Novelty September 2019

[2]Code source from https://scikit-learn.org/

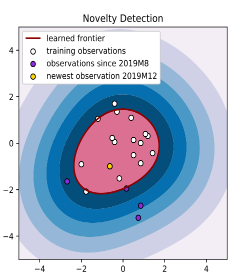


[October]

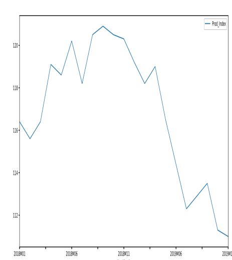


[November]

Figure 8: October and November Novelties 2019



[December]



[Industrial Production]

Figure 9: December Value and Industrial Production Index

The testing sample has 5 observations. Each month, starting August 2019, I draw a new graph, to find out if the newest observation lies outside the learned frontier, as it should. This is the observation in yellow. Testing observations are in purple. As one could see from the graph, August was classifed as the worst month while September and December were not even classified as novelties. According to the current classification, one would expect to see an improvement in the hard data. Hard data in fact went further down into November 2019, while Export Soft Indicator turned into a worst level in November 2019. The current sample is too tainted to be able to classify data correctly.

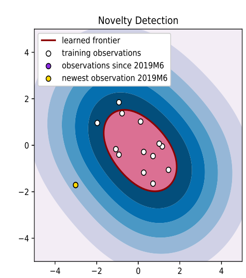
4.2 Training Sample April 2018 to May 2019

I use as training sample April 2018 to May 2019, corresponding to the first breaks in industry index and last break in export soft indicator. I would liket to know how sensitive is the classification mechanism to this tainted value. The test sample is from June 2019 to December 2019.

In this case only two observations out of 13 in the training sample were misclassified, corresponding to an error of 0.1 as set above. They were very close to the frontier, in the dark blue area (see graphs below). In general the learned frontier has a different shape, where values classified as novelties have to be in the left down corner, where both values for Recent Trend and Exports are negative. Recent Trend data, therefore domestic causes, are the main driving force during this location shift, as most negative values come from Recent Trend data.

Each month, starting June 2019, I draw a new graph, to find out if the newest observation lies outside the learned frontier. This is the observation in yellow. Testing observations are in purple. Below I document the evolution of the classification outcome from June to December 2019.

[June - first]



[July]

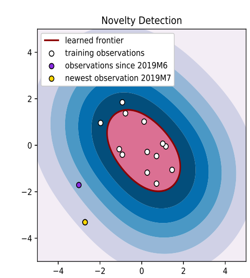


Figure 10: Novelties in June and July 2019

Below it is October graph. The worst value was so far in October 2019, classified outside the gray circle, the outermost area of the distribution (see Figure 11). It could be the turning point towards an improved output. Both

[3]Code source from https://scikit-learn.org/

external and internal factors were at their worst levels, with external factors worse off.

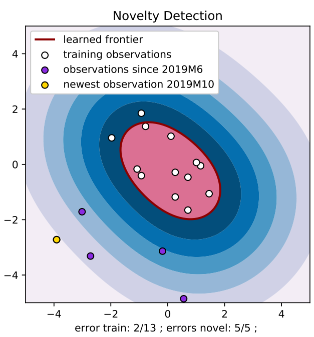
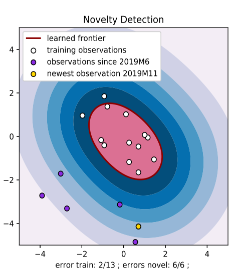


Figure 11: October Novelty

Graphs below (see Figure 12) are for the last two months, November and December 2019.

[November] 

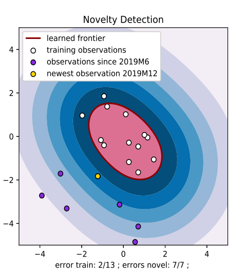
[December - the last] 

Figure 12: Novelties in November and December 2019

Values improved in November and December. December, even though closer to frontier, remained in the second blue region. It is still a novelty. Based on this classification method, months of hardship are still ahead of us.

5 Conclusion

In this paper One-Class Support Vector Machine uses a combination of soft indicators to describe the evolution of production index during its most recent location shift that started in March 2018. This was a result of both internal factors (increased wages in Romania) and external factors (technological changes along the value-added chain coming from Germany and France). The goal is to help policy makers understand the dynamics of location shifts. This time the switch to a new equilibrium goes through a period of hardship. In fact

Romanian industry is moving to a new equilibrium output level with increased productivity, higher wages, and increased export competitiveness, hand in hand with Euro Area industry sector. Ideally one would like to better understand the phases of this process of equilibrium formation. According to Castle et. al (op.cit.) there is normally an extreme value, a 'worst point' or a 'best point', the one that marks the turning point in a location shift. So far, October 2019 was classified as the worst point. However, it is too early to tell more about the evolution of this sector at this point.

A second benefit of this paper is that it has a first look at how animal spirit (as captured by survey indicators) evolves during this equilibrium formation process. There is a wealth of survey indicators. They are faster than hard indicators, and therefore their power could be harnessed to better understand the process of equilibrium formation. One conclusion is that animal spirit is not always ahead of supply changes. They seem to be inter-related, with hard-core changes reflected in the 'mood' of business participants, and then going back into the hard-core indicator. This relationship is therefore rather circular, similar to that of inter-temporal co-breaking. Scattterplots captured this behavior, showing often 'clouds' and 'distances' in the plots. One-Class SVM exploits these distances in finding the end/the start of a location shift.

At the stage, the current location shift is not over. The method described above is to be continued in classifying observations in the coming months.

References

Castle, J., Clements, M., and Hendry, D. 2016. *Overview of Forecasting Facing Breaks*, Journal of Business Cycle, 12:3-23.

Clements, M. and Hendry, D. 2001. *Forecasting Non-stationary Economic Time Series*. MIT Press.

Farmer, R. and Woodford, M. 1997. *Self-fulfilling Prophecies and the Business Cycle*. Macroeconomic Dynamics I: 740-69.

Farmer, R. 1999. *Macroeconomics of Self-fulfilling Prophecies*. MIT Press.

Hendry, D. F. 1995. *Dynamic Econometrics*. Oxford: Oxford University Press.

Rappoport, P. and Reichlin, L. 1989. *Segmented Trends and Non-Stationary Time Series*. Economic Journal 99: 168-177.

Schoelkopf, B., Williamson, R., Smola, A., Shawe-Taylor, J., Platt, J. 2000. *Support Vector Method for Novelty Detection. Advances in Neural Information Processing Systems: Proceedings of the 1999 Conference*. http://papers.nips.cc/paper/1723-support-vector-method-for-novelty-detection.pdf.

Smola, A. and Schoelkopf, B. 2004. *A Tutorial on Support Vector Regression*. Statistics and Computing 14: 199–222.