

# Report on Bachelor / Master Thesis

Institute of Economic Studies, Faculty of Social Sciences, Charles University in Prague

<b>Student:</b>	<b>Jan Malecha</b>
<b>Advisor:</b>	<b>Petr Pleticha, MSc.</b>
<b>Title of the thesis:</b>	<b>Innovation Indicator Analysis in the European Union: A Machine Learning Approach</b>

## **OVERALL ASSESSMENT** (provided in English, Czech, or Slovak):

*Please provide your assessment of each of the following four categories, summary and suggested questions for the discussion. The minimum length of the report is 300 words.*

### **Contribution**

Although it might not have been the main original objective, the author rather credibly demonstrates the weaknesses of the European Innovation Survey (EIS) data and of the simplified original attempts to use such data for a classification of countries into four distinct categories based on a very simple concept such as the Synthetic Innovation Index (SII). Both the implications of his literature review as well as his own results show the rather ad-hoc nature of this EU attempt to evaluate innovation performance.

The author also describes an interesting relationship between the identified clusters and the sequencing of EU enlargement – and the apparent deviation of the 4 southern EU members from this pattern. That's an interesting conclusion that might have deserved additional analysis.

The results of the attempt to identify the determinants of productivity of labour are a bit less convincing (because of the nature of the data as well as methodology), but they are still at a level adequate for an undergraduate student.

### **Methods**

The author analyzes the European Innovation Survey (EIS) data with the use of two categories of methods: (i) hierarchical clustering by means of which he tests the logic of the original classification of all EU countries into four categories, (ii) regression analysis which attempts to test the relationship between productivity of labour and the 27 variables available from the EIS, he also complements the regression analysis with lasso by means of which he attempts to identify the truly relevant variables. The application of hierarchical clustering can be criticized from the same point of view that the author describes in his literature review. If – as Edquist et al. (2018) argue - the Synthetic Innovation Index (SII) can be criticized for mixing the input and output characteristics (p. 7), doesn't the author's decision to use all the available variables in his attempt to clustering lead to a very similar problem? And wouldn't it therefore make sense to differentiate between these variables in the attempts at clustering and tests clustering with the subsets of the variables too?

The design of the regression specification is a bit mechanical, it is not derived from any micro-founded model – but that is quite common within the class of publications which rely on machine learning approaches. Still, would not it be logical to test the interaction effects too? There may be some interactions between effects of R&D expenditures on education and the number of graduates or the quality of domestic R&D sector (measured by patents and publications). Last but not least, it might have been interesting to use also an alternative dependent variable, e.g. TFP.

### **Literature**

The literature review is relatively brief. The parts of the review which focus on the European Innovation Survey (EIS) seem to be quite OK and in spite of their brevity they very well characterize the EIS and its main problems identified by previous authors. However, it seemed to me that the author did not always fully reflect the findings from the literature review in the design of his own research – the aforementioned case of the criticism of the EIS by Edquist et. al. (2018).

The discussion of previous research on the determinants of productivity could have been perhaps a bit more extensive, it is a widely researched topic. I would have loved to see there a more explicit debate on the proper measurement of productivity (productivity of labour vs. TFP) and also e.g. the role of

# Report on Bachelor / Master Thesis

Institute of Economic Studies, Faculty of Social Sciences, Charles University in Prague

<b>Student:</b>	<b>Jan Malecha</b>
<b>Advisor:</b>	<b>Petr Pleticha, MSc.</b>
<b>Title of the thesis:</b>	<b>Innovation Indicator Analysis in the European Union: A Machine Learning Approach</b>

differences in specialization patterns and of capital endowment of labour in the observed differences in the “productivity of labour”.

## Manuscript form

The style of the manuscript was quite appropriate for a bachelor thesis. I would perhaps only recommend that the author omits the unnecessary description of the logic of the pooled OLS (p.15-16). There were relatively few typos left in the manuscript.

The graphical form of the manuscript was at a very high level, the author used graphical tools (graphical output from clustering procedures) convincingly and clearly.

## Summary and suggested questions for the discussion during the defense

The submitted thesis meets the criteria for bachelor theses at the IES and I recommend it for defense.

Suggested questions:

1. Why did not you try to build on the conclusions of Edquist et al. (2018) and did not differentiate between input and output indicators available in the EIS? Were there any technical or practical reasons for this decision?
2. The author concludes the thesis by suggesting that it might be interesting to use Global Innovation Index data (rather than the EIS data). Is this dataset more consistent and would it therefore reduce some of the objections raised by Schibany & Streicher (2008) and Edquist et. Al (2018) against the EIS and SII?

## **SUMMARY OF POINTS AWARDED** (for details, see below):

<b>CATEGORY</b>	<b>POINTS</b>
<i>Contribution</i> (max. 30 points)	22
<i>Methods</i> (max. 30 points)	20
<i>Literature</i> (max. 20 points)	16
<i>Manuscript Form</i> (max. 20 points)	20
<b>TOTAL POINTS</b> (max. 100 points)	<b>78</b>
<b>GRADE</b> (A – B – C – D – E – F)	<b>C</b>

**NAME OF THE REFEREE:** *Vilém Semerák, Ph.D.*

**DATE OF EVALUATION:** *May 30th, 2019*

---

**Referee Signature**

**EXPLANATION OF CATEGORIES AND SCALE:**

**CONTRIBUTION:** *The author presents original ideas on the topic demonstrating critical thinking and ability to draw conclusions based on the knowledge of relevant theory and empirics. There is a distinct value added of the thesis.*

<i>Strong</i>	<i>Average</i>	<i>Weak</i>
30	15	0

**METHODS:** *The tools used are relevant to the research question being investigated, and adequate to the author's level of studies. The thesis topic is comprehensively analyzed.*

<i>Strong</i>	<i>Average</i>	<i>Weak</i>
30	15	0

**LITERATURE REVIEW:** *The thesis demonstrates author's full understanding and command of recent literature. The author quotes relevant literature in a proper way.*

<i>Strong</i>	<i>Average</i>	<i>Weak</i>
20	10	0

**MANUSCRIPT FORM:** *The thesis is well structured. The student uses appropriate language and style, including academic format for graphs and tables. The text effectively refers to graphs and tables and disposes with a complete bibliography.*

<i>Strong</i>	<i>Average</i>	<i>Weak</i>
20	10	0

**Overall grading:**

TOTAL	GRADE
91 – 100	A
81 - 90	B
71 - 80	C
61 – 70	D
51 – 60	E
0 – 50	F