GREThA-UMR CNRS 5113
Université de Bordeaux
av.Léon Duguit
33608 Pessac cedex (France)

Prof. Kresimir Zigic
CERGE-EI
Charles University
Prague

Bordeaux,

10/05/2016

Objet: Report on the doctoral thesis submitted by Oleg Sidorkin

The thesis consists of a collection of three distinct papers, two of which in co-authorship. The first two papers deal with topics within the field of the economics of innovation. The last paper is dedicated to an altogether different topic, namely the economics of corruption, which belongs to what is currently referred to as political economy. Despite this difference in contents, the last paper shares with the former two several methodological traits, namely a quantitative approach based upon firm level, questionnaire-based micro-data; and some degree of sophistication in the econometric treatment. One can sense that the candidate is more attracted by the methods he uses, than the substantive contents of his enquiries.

As a general comment on style and accuracy of the thesis, I recommend an accurate revision and editing, possibly with the help of a native English speaker. Paper 1 contains a large number of colloquial expressions not be used in a PhD dissertation. Besides sounding out of place, they add to the general feeling of untidiness in the presentation of the main hypotheses and of the data used. Paper 2 requires restructuring (more on this below). Paper 3 reads like to most accomplished piece of work, but still some restricting would help.

As I am not at all competent in the last paper's topics, I will devote most of this review to the first two, especially as far as substantive comments are concerned.

Paper 1 deals with a classic topic in the economics of innovation, namely the relationship between innovation and survival or growth at the firm level. It does so by means of a cross-section analysis of survey data matched to financial information, for a sample of firms from several Eastern European countries, plus Turkey. The core hypotheses to be tested can be summarized as follows:

1. Are innovative firms are more likely to survive (and grow)?
2. Is the innovation-survival relationship non-linear, with the most innovative firms being less likely to survive than the moderately innovative ones?

The contribution is extremely valuable in that it cast light on a number of countries most often ignored in the recent literature on innovation, and it does so with captivating micro-data. The econometrics is state-of-the-art, although some choices look questionable and should be better motivated.

The data used by the candidate refer to survival through the last recession, which pushes him to present his findings as specific to firm survival during a recession phase, as well as to comment/explain them is such a light. I would recommend him to be more cautious on this point, as he has data only for the recession phase, so he cannot test for whether his findings are specific to it or not (unless he finds the way to exploit cross-country variability in his data, to the extent that different countries have been differently affected by the crisis).

The presentation of the data set is quite untidy and the use of casual language I mentioned above does not help. (Even the numeration of tables is confusing: Why relegating all tables in an Appendix? Some tables include "A" in the number, some do not; table 1.5 and 1.6 come after 1.7…). Therefore, I am not 100% sure of whether some of the remarks that follow are appropriate, being the case that they possibly originate from a misunderstanding on my part. If this is the case, the candidate should ignore them, but be aware that this reinforces the case for revising and re-editing the paper.

- The BEEPS survey is used both in paper 1 and paper 3. Both papers are written as stand-alone chapters, following a format which is clearly for publication on a journal. Still, given the importance of the source and the fact that a PhD dissertation has no size limitations, I would suggest to create an Appendix to the thesis (not to each paper), entirely dedicated to presenting the sources (BEEPS above all, but also the FCS data matched to it in paper 1 and MOI used in paper 2). Many of the requests for further clarification I put forward below can be accommodated in this way without changing too much each paper. Besides, the candidate ought to be aware that most journals now accommodate for online supplementary material, which means that the appendix I propose could be easily recycled to that purpose.
- There is no information whatsoever in the chapter on the sampling scheme followed by BEEPS and its representativeness. Besides, I had a very hard time to understand how the candidate reached the final sample for his econometric exercise. In this respect, he mentions 3 samples, respectively of 1489, 1247 and 582 observations. The first one is unacceptable, as it is used only for the INNOVATION regression (column 1 Table 1.3), while the core regression (the one with SURVIVAL as dependent variable) is clearly based on 1247. THAT is the sample, better not messing around (as for 582, it is ok to mention it, as it concerns a dataset which is nested in the main one). I have also some doubt on whether it is appropriate to calculate the INNOVexcess and INNOVcautious dummies on a larger sample than the one used in the main regression, as the position of the firms in the Kernel distribution is affected by several firms whose survival rate we do not know
- A key substantive issue is the definition of the INNOVATION variables. It is not clear whether:
    - It concerns novelty for the firm as opposed to novelty for the market (a classic distinction in the CIS survey, which I believe BEEPS considers, too)
    - It concerns product or process innovation. I ask this because the only information I get from the paper is that it measures the shares (in discrete units from 0 to 100) of sales due to "innovated" products. Now, innovated is the past tense of the verb "to innovate; as an adjective, it does not exist in the English language. [A quick search on Ngram Google will reveal it is used rarely; and a causal check of Google Scholar will show that publications that use it come from non-English speakers and are barely cited]. Either we are talking of sales of new products or innovative products (in which case I tend to think we are talking of product innovation), or products which incorporate some innovation (which may be a process innovation).

  The issue is crucial for the interpretation of the results. In one case, the result is very specific, and more convincing: if the new products are profitable only if sold with a price premium,

consumers may refrain from paying it during recession. In the other case, the result is less specific and less clear, too, as one would expect that firms having invested in process innovation may meet the consumers' taste for cheaper products during recession. The new-to-the-firm vs new-to-the-market distinction matters, too, as the former may simply indicated a firm's effort to catch up, while the second one has to do with leadership and risk (more in line with the references to Schumpeter used by the candidate; otherwise, the best references to Schumpeter would be those in which the latter explained how imitators may introduce their own innovations when it is too late).

- Concerning the model specification and the way the results are interpreted, I wonder whether it is really necessary to run the INNOV regression, create the INNOVexcess and INNOVcautious dummies and then use them in the main regression. Wouldn't inserting a quadratic term make the same trick, provided that controls are the same in the INNOV and the SURVIVAL regression? Maybe I am wrong, but the candidate should work much harder to make the case for the specification of his choice. Notice also that the robustness checks (changes of thresholds in the kernel distribution for the choice of INNOVexcess and INNOVcautious) do not go very much in the direction of reassuring the reader. Most of them fail to deliver: significance disappears (not only when the number of observations beyond the threshold decreases, but also when it increases), and, worse, the size of the coefficients change considerably (all together, it makes one wonder whether the results are dictated by a few outliers only).
- ZINB for the INNOV regression is a very odd choice, and quite poorly motivated. INNOV is not a count variable, its discrete nature being simply an artifact due to having dropped the decimals. There are selection regressions (starting from Tobit) that are expressly thought for asymmetrically distributed continuous variables: why not using them? What are the advantages of using ZINB? Is there any literature that supports this choice?
- Are equations from (1.1) to (1.3) and the related digression on ZINB really necessary? It is textbook econometrics with no specific information on its use in the paper (and the notation is messy)
- I am sympathetic with using OLS instead of Logit/Probit, when possible, but the candidate has a problem in convincing me (despite the very long, too long explanation of his choice):
    - The SURVIVAL dependent variable is very asymmetrically distributed. If I am not wrong, 93% of observations have value 1. In this case, authors such as Paul Allisson or J.Scott Long suggest caution.
    - The candidate provides scant information on the distribution of predictions: the linear probability model (i.e. OLS) approximates well the S-shaped distribution of Prob(y=1) only around the inflexion point, hence the recommendation to use it only if most predictions fall in in between 0.2 and 0.8. But it seems that many predictions are over 1 (we are told nothing on <0 predictions
    - Even if one uses OLS, it is good practice to report Logit estimates in an Appendix, and to show that results are similar. Not doing so raises suspicions on the robustness of the exercise
- Table 1.6 is a mess. It reports too much, contradictory information. The bottom line indicates the total sample size of three waves of FCS, but actually only the second wave matters (and one has to dig into the text in order to understand it). And none of the three totals match the number of observations indicated under the table (1247), which hover coincides with the sample used in the regressions. Even for the second wave, the sum of observations with SURVIVE=1 and =0 is 1184, not 1247. What's going on?

Moving to paper 2, I found it much less interesting than paper 1. Problems of endogeneity, although present also in paper 1, here are huge and not treated at all (albeit the author rightly admit them). The research question is also much less stimulating (although this is a matter of personal taste on my part). More generally, one has the sense of a hasty piece of work. It happens that one paper is quickly drafted

to finish the PhD, but I think that, even without asking the candidate to fully develop it right now, something can be done to make it more readable. First, the author should split what is now section 2.1 into a proper introduction (one that summarizes the paper) and a literature review. The two are now conflated into one. Besides, as it is clear that the hypotheses to test for this paper come straight out of the literature review, this should be more linear:

- first the review of the literature on managerial practices and innovation (what is now split between pages 29-30 and the second half og page 31 + page 32);
- then the derivation of the hypotheses to be tested (which is now inserted like a wedge in between the two bits of literature review, in the first half of page 31).

As with paper 1, the data structure and properties are poorly explained. For example, we know that fata come from the EBRD (which dataset? Is it accessible) and MOI, at some point a "match" between the two is mentioned, but we do not know how the match was made (e.g. through a unique identifier or just by the company's name) and which variables come from which source. Why are financial data missing for 36% of companies?

The way the answers to MOI questions are coded (scores assigned, normalization, sum, further normalization etc) is very complex and not motivated. Why is it necessary to normalize and then sum the answers (also in light that there is just one answer, or set of answer, that really matters)? Conditional on aggregating the answers, why not using more common techniques for treating survey data, such as principal component analysis? How sensitive are the econometric results to this type of data treatment?

The hypothesis that higher scores correspond to best practices is not immediately convincing for all managerial practices. Can the candidate elaborate more on it? Is it possible that the lack of significant results for several covariates is explained by problems with this hypothesis?

Coming to the last paper, I admit once again my lack of expertise in the field, so I will limit myself to a few comments on structure, clarity and interpretation

On structure, I find it odd to see first some qualitative analysis of data, which anticipate most of the theoretical arguments; then a formal model, which systematize the same arguments; than an econometric test of the formal model. I think it would be more rational to go first for the theoretical model, then for all data analysis. Many of the hypotheses upon which the theory rests (such as the absence of reverse causality, from corruption to non-renewal of the nomination) ought to be discussed along with the theoretical model, and not as after-thoughts at the moment of the interpretation of the econometric results. Eventually, the discussion of econometric results ought to test more clearly such important hypotheses (this is partly done in table 3.6, although the Time covariates suddenly disappears, which looks suspicious).

I am confident that the candidate will review his thesis according to the comments I have provided, although I guess it will take some time. Conditional on this, I judge that thesis satisfies formal and content requirements for a PhD thesis in economics, and I recommend for its defence to take place.