Referee report on the manuscript "Where do new firms locate ? The effects of Agglomeration on the Formation and Scale of Operations of New Firms in Punjab"

The authors analyze whether local industrial characteristics influence the location choice of firms in the case of new firms creation. The analysis is done on a sample of Pakistanis firms in 2008, using the Directory of Industrial firms for Pakistan. The authors develop an estimated equation from a theoretical model by Soubeyran and Thisse (1998). They estimate the presence of urbanization and localization economies, as the impact of incumbent firms on the number of newly created firms and on their size.

The paper is nicely written and this should be emphasized. In particular, I appreciate the fact that the authors take time to explain some of the intuitions behind the theoretical and empirical analysis. This is not always the case, and this should be of course kept as an attribute of the revised paper. There are a number of important issues in the manuscript, which should be addressed in order to strengthen the estimation, the contribution and the way the paper is presented. I list them below, in the order of the paper's sections.

Introduction, question of the paper

Although the question addressed in the paper is interesting, I think that the introduction misleads the reader on the exact issue which is analyzed, for the following reasons.

- From the first sentence of the introduction, the reader is brought to think about the benefits of the formation of new firms. There is a difference between understanding whether new firms do appear, and why, and understanding where do new firms appear, conditionally on being created. The paper addresses the second question and should stick to this presentation.
- The second sentence is also very misleading to me. The "well established socio-economic benefits " are not defined (until the next section), and come as if we had been discussing the issue for a long time. This relates to the direct effects of the creation of new firms on individuals of the neighborhood.
- Then, the theoretical elements come into the discussion, which is nice. However the authors cite Marshall and do not cite Duranton and Puga, which have written a very nice and relatively recent literature review on the theoretical explanations of agglomeration economies.
- Then, the Jacob-type of agglomeration economies come into the discussion, which are a type of agglomeration economies and could be brought forth somewhere else than in the introduction. The broad picture is missing.
- Other empirical papers are cited, however we still do not know what the current paper does.

- The paper addresses the question of agglomeration externalities, with a focus on learning externalities. Sometimes the learning externalities focus disappears in the paper.

Existing literature and references

There are several issues about the cited (and not cited) papers in the manuscript.

First, in my view there is a lack of references to the existing literature in agglomeration, trade and location economics. Also, the cited references are sometimes misplaced. For exemple, Otsuka (2008) comes in the first page and appears as a leading paper, while there could be other papers placed before, linked to the current topic.

Sorenson (2000) is a sociology paper. It should be said, both because the reference comes from outside the IO and trade literature, and because it is interesting to say that people in sociology study these types of questions. However the citation should maybe not figure as a main reference in the manuscript.

Papers on location choices in the trade literature are for example *Crozet M., T. Mayer , J.-L. Mucchielli 2004. "How Do Firms Agglomerate? A Study of FDI in France" Regional Science and Urban Economics Vol. 34 (1), January: 27-54.).*
As a reader, I would like to see citations of the frontier of this type of empirical research, and also, more specifically, this type of research focused on learning externalities. The authors should explain the contribution of the location choice literature, and also the learning externalities literature.

As a general comment I think the authors should take a step back from the exact question addressed in their paper, so as to be able to present their paper as part of the agglomeration and trade literature. Again, the current citations are not wrong, however they are to focused on "new firm formation". It would indeed by an interesting idea to explain to the reader the difference between new firm formation and location choices of firms.

Finally some references do not figure in the Ref. section. Duranton and Puga for example. Some references do figure in this section and should maybe not. Where does the paper on contracting and efficiency in the surgical sector come into the discussion ?

Theoretical model

The theoretical model by Soubeyran and Thisse is really interesting to read. However, the reader does not fully understand the necessity to have section 3 explain large parts of the model, when the empirical specification is based on equation 10. Instead of erasing the theoretical part, I would like to see the model better explained, i.e. to have the entire section 3 oriented towards obtaining a nice empirical specification.

Data

The authors use data on Pakistanese firms. Some questions arise. Why only the years 2006 and 2010 ? The first paragraph in section 4 is not that clear. Is the database exhaustive ? The sentence "We have used the DOI 2010 to measure the arrival of firms in 2008" is not clear. If the reverse causality issue is the main explanation, then the arguments should be made the other way around.

More information should be given on the dataset. Are we sure that these are firm creations ? Can't these be firm movements ? Before showing us data on the potential urbanization and localization issues, we would like to see comparisons across years, for example. Maybe references to other countries. Again, here the reader should understand why the "firm creation" issue is specific and different from an increase in employment, for example, or a multinational firm's location choice. I could be interested in asking whether the firms which increase their employment in the considered period, are more specifically located in the agglomerated areas. Is this question different from the question addressed by the authors ?

Specification

The two specifications are very interesting questions. However, they need to be placed in the existing literature. For instance, I don't see where the demand effect is addressed, nor the competition effect. Choosing an oligopolistic model to explain location is a very good idea, and we should see the consequences of this in the specification and in the text.

Why not analyze the location choice at the firm level ? The dataset is at the firm-level.

The authors should explain which is the variability which explains the estimation. The explained variable is the total number of new firms in industry i and district d in a single year. Why do we have only one year ? Using the panel could allow to control for all the regional characteristics. Is the only variability here, for a given district, among the different industries, and for a given industry, among the different districts ? The authors should be clearer about the interpretations they want to push forward.

Minor remark: The explanation of the two estimated equation begin by presenting the error terms. This is not adequate.

The expression "socio economic factors" is repeatedly used. This should be avoided, a well as for other often repeated "ready to go" expressions .

The results section is nicely explained. However it could be a little longer, with an analysis which could take a step back. Comparing with other existing studies is interesting, but I don't see papers in the trade literature here. Also, results should be discussed differently. Selection effect is very important. Say more on this. Say more on

oligopolistic tupe of behavior also. How does the model by Soubeyran and Thisse intervene in shaping these results ? Is it necessary to have the specifications taken out of this model specifically ? This is really important to add.

There seem to be a problem with section 6.1. Why is the robustness section so short ? Also, checking the same estimations without controls does not really represent robustness checks.