Comments on "Where Do New Firms Locate? The Effects of Agglomeration on the Formation and Scale of Operations of New Firms in Punjab"

General

Section 1 and 2 are well written and give a good overview of the localization/agglomeration issues as they appear in recent literature. However, there is no link with section 3 where a strict agglomeration concept is chosen from Soubeyran & Thisse without any further motivation.

Theoretical framework

page 5-6 raises many questions that need to be cleared:

K_d is introduced as the initial level of knowledge in district d (line 6). One sentence further (line 7) the same symbol K_d is introduced as capital for which interest rate r is charged. In the latter I assume it should read k assuming that this k is the same in every district? Further confusing come in when later it is said that capital $k(q_d)$ "is constant across districts" but apparently it is not fixed costs but only variable costs depend on output

In (1) symbol l is introduced as "worker's knowledge base", but then, what is $l(K_d)$? The worker's knowledge base as a function of the initial level of knowledge in district d? And why is it multiplied with output q_d and the wage w_d, apparently to represent labor cost? I suspect that l should be l_d here meaning the number of workers hired in district d which is a negative function of K_d, presumably because with a higher knowledge base you need less labor? But then it should read $l(K_d, q_d)$ because, just as $k(q_d)$, the required number of workers must be some function of output.

From equation (2) it is clear that maximizing profits is the same as minimizing costs. These costs go down with K_d by definition because $l' < 0$ by definition. Therefore equation (3) is confusing and can be deleted. Apart from that, define what you mean with the hat (^) symbol here.

Equation (4) is confusing: what is K' (the first derivative to what?). Equation (5) is self-evident and can be deleted. Equation (6) does not tell what λ is.

Equation (7) should be deleted because defining the number of firms in (8) as total labor force divided by labor demand of one firm is clear enough. What is not clear, however, is why every firm is of the same size. This is an implicit assumption in (8) and needs to be addressed.

"Output produced by firms is the same across districts" sounds like every district is of the same size, you probably mean identical firm size across all districts, which as far as I can see cannot be derived here but is an assumption.

In equation (9) what is (hat) $q(I)$? Your previous notation to indicate output by district is q_d. Better use Q to indicate total output of all districts together which I believe you mean here. The symbol v is said to be increasing here but it does not tell what it is. Better just delete (9) and say that you assume all labor supply is used in each region. This directly gives the number of firms in each district to be linear proportional to labor supply by district. In the end, the only qualification that comes out of page 5-6 in equation (10) is that n_d will be less that proportional to L_d because a higher stock of knowledge means more productive labor. For that, you might even consider to delete all equations because they do not add to any further understanding (redundant math exposure).

Some more theory (and maybe equations) is needed here, however, to justify the assumption of all firms being of the same size. This is normally an outcome of monopolistic competition models, which I expect Soubeyran & Thisse are also using.

Data sources and descriptive statistics

Explain why and how you use the DOI 2010 to measure the arrival of firms in 2008 only. Why not for 2010? If that is possible, you can also make a dataset of arrivals for each individual year? In addition, does arrival mean a firm coming from somewhere else or is it a startup firm? Also interesting is if it is an independent firm. Maybe many of these arrivals could be branches or establishments of the same larger company (would you consider every opening of a new McDonald as an arrival of a new firm?).

The paper does not mention how many districts appear in the dataset which makes it unclear to interpret the results and degrees of freedom of the model. Please add a table with the nr of observations by district. Also, how large are the districts? When they are small there could be important spillover effects of local knowledge bases between them and spatial autocorrelation effects.

Empirical specification

Equation (11) and (12) follow quite nicely from the literature discussed on page 1-4. Again this indicates that section 3 as a whole can better be deleted from the paper because it is not related in any way to the empirical specification.

What are "sub provincial effects" and why are they added? Is this something else than district? How many sub provinces are there?

For a measure of localization for industry i in district d, simply total employment have been taken. This is not localization but just size. Your model then predicts that more arrivals will happen in large districts. Instead, you need a localization indicator λ_d like for example the regional component of the sector share: local share of employment in – say – the footwear sector divided by the same at the national/provincial level). Then, if $\lambda_d > 1$, district d is relatively more specialized in footwear. See Combes, Mayer & Thisse (2008) chapter 10 for an elaborate discussion on localization and specialization measures.

For a measure of urbanization the same problem appears. Here simply total employment is taken which again is only measuring size. Density, or the share of population of the largest city(ies) in the district could be taken.

Results

Given the above, it comes as no surprise that everything is very significant. I will not further discuss the results because I believe they do not have any meaning.

References

Combes, P., T. Mayer & J. Thisse (2008) *Economic Geography, The integration of regions and nations*, Princeton: Princeton University Press