GitHub users in Spain: an geospatial analysis

JJ Merelo

18/04/2015

Abstract

Measuring and ranking the free software developers in a particular geographical space is a way of knowing the existing community and also allows assessing the impact of certain policies in the dynamics of such a community. Besides, it is interesting to try and find out why there are differences from one place to the next and how these differences evolve with time. In this paper, our main interest is to measure and rank the community of free software developers in Spain and also check its geographical distribution. This paper measures differences by province, providing a classification of provinces according to the number and type of developers present in each place.

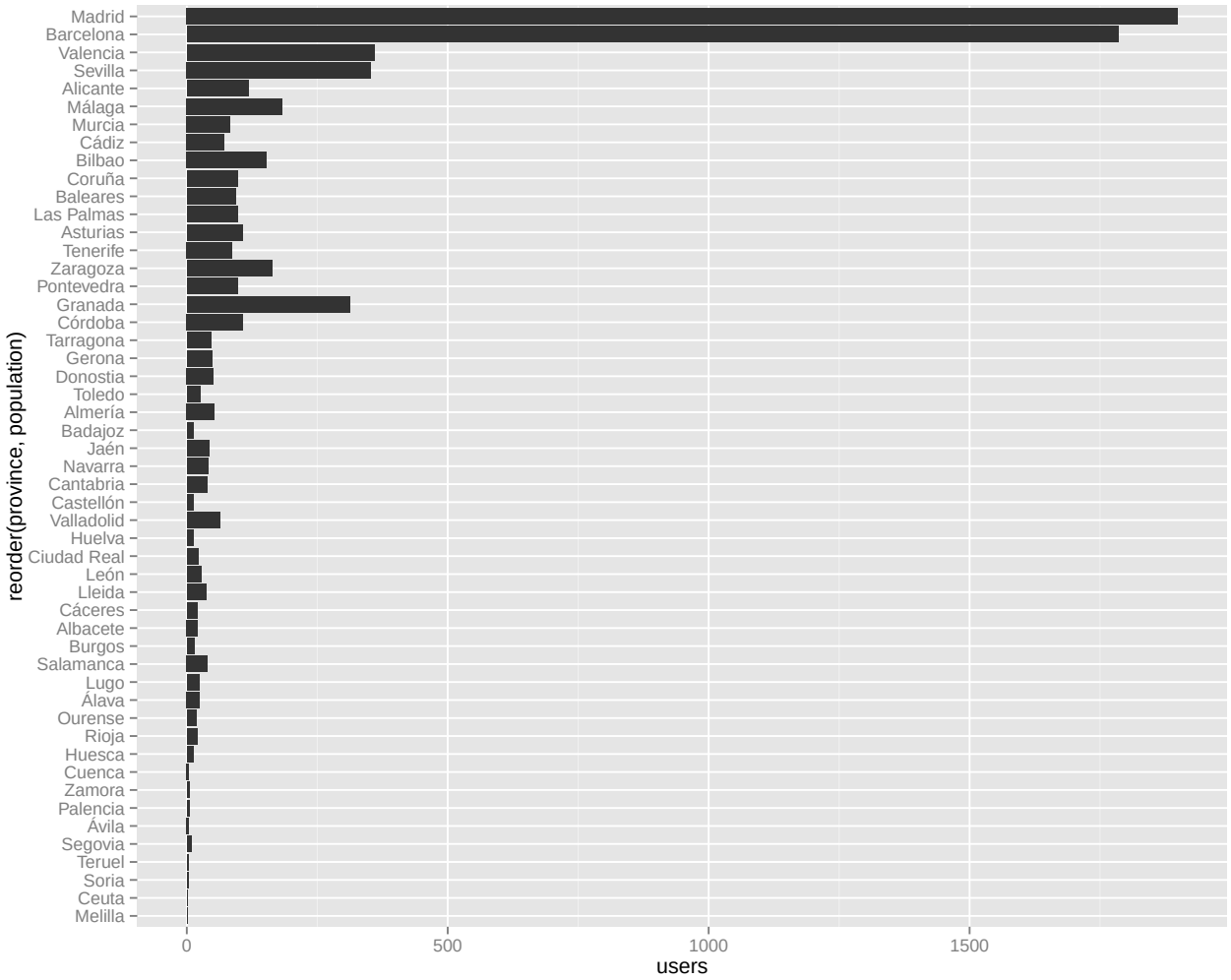
Introduction and methodology

The initial motive behind this paper was to check the health of the free software developer community in Spain. With that target in mind, we elaborated some initial national rankings which were published Merelo-Guervos et al. (2015). The main problem with those rankings is that in many cases and specially in the big cities, there was no attempt to exhaustively search all active users. Since the GitHub search API just returns 1000 results and, in the case of Madrid and Barcelona, there were way more than that, the script that downloads user data had to be modified so that, through partition of the search space, it was able to perform searches that returned less than 1000 results until all users were covered. It still does not cover users that *do not* declare their city/province in their profile, or use provincial towns thar are not explicitly searched. In one case, Guadalajara, it was impossible to make out which users where actually from Guadalajara, Spain and not from Guadalajara, Mexico and thus was explicitly excluded. In general, it can be said that all users that *declare* their province or provincial capital are included, although the quantity of those that are there and do not do it is unknown, and hopefully uniform for all provinces involved.

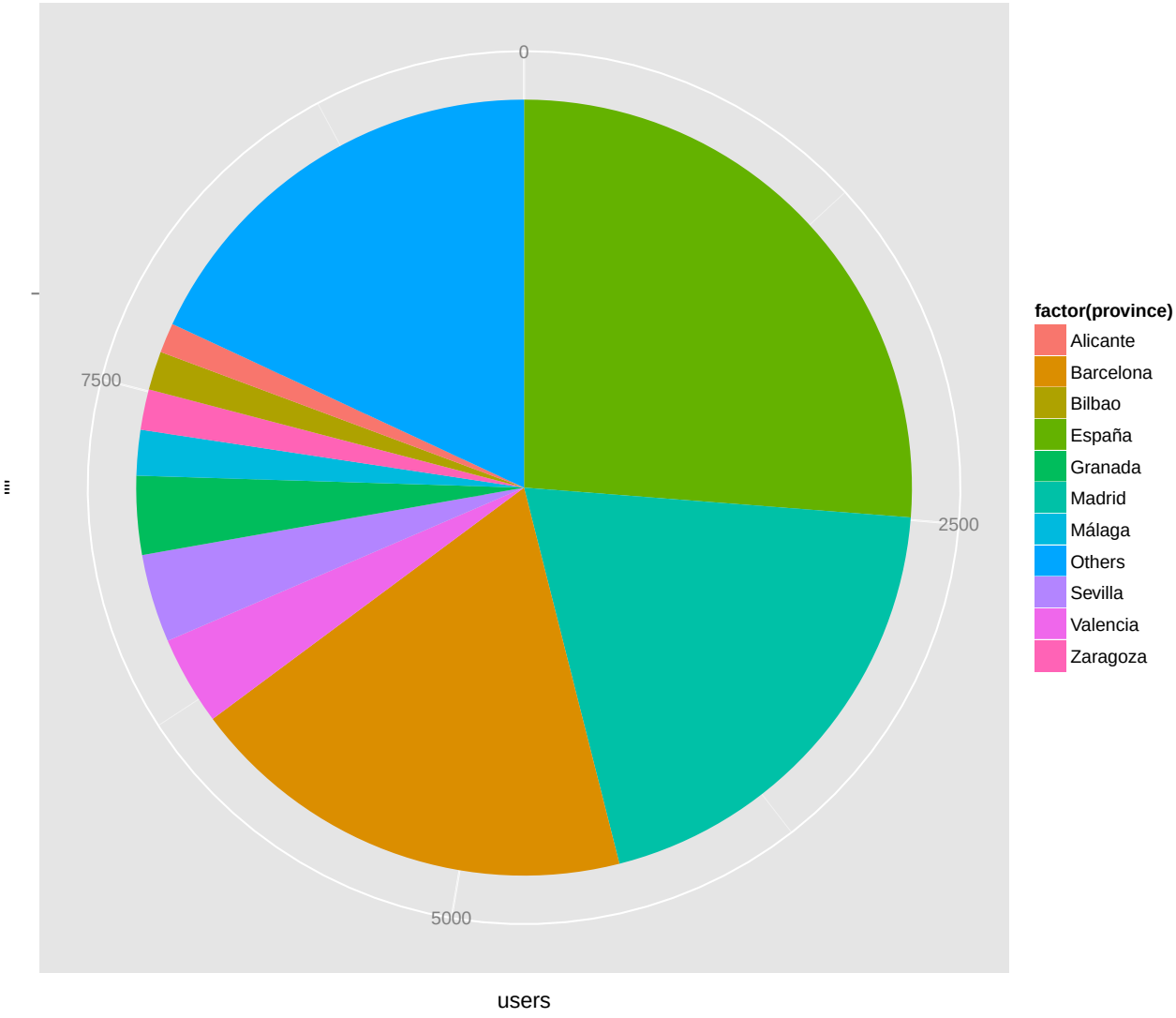
In a previous version of this paper Merelo (2015), we had not been able to obtain all the results for the whole country, that is, users that declare their country (plus a city or town or community that is not searched otherwise, that is, excluding the provincial town). However, in this paper all users that declare Spain (in several versions) as their place of residence have been retrieved and additional analysis that involves them can be performed. This will be done next

Results

After downloading all users, scraping was performed over the user profiles to extract the following information: number of followers, stars given by the user, stars given to the projects in which the user participates, and raw number of users. This number of users in each province is shown below. Obviously, the provinces with the biggest population do have the bigger number of users.

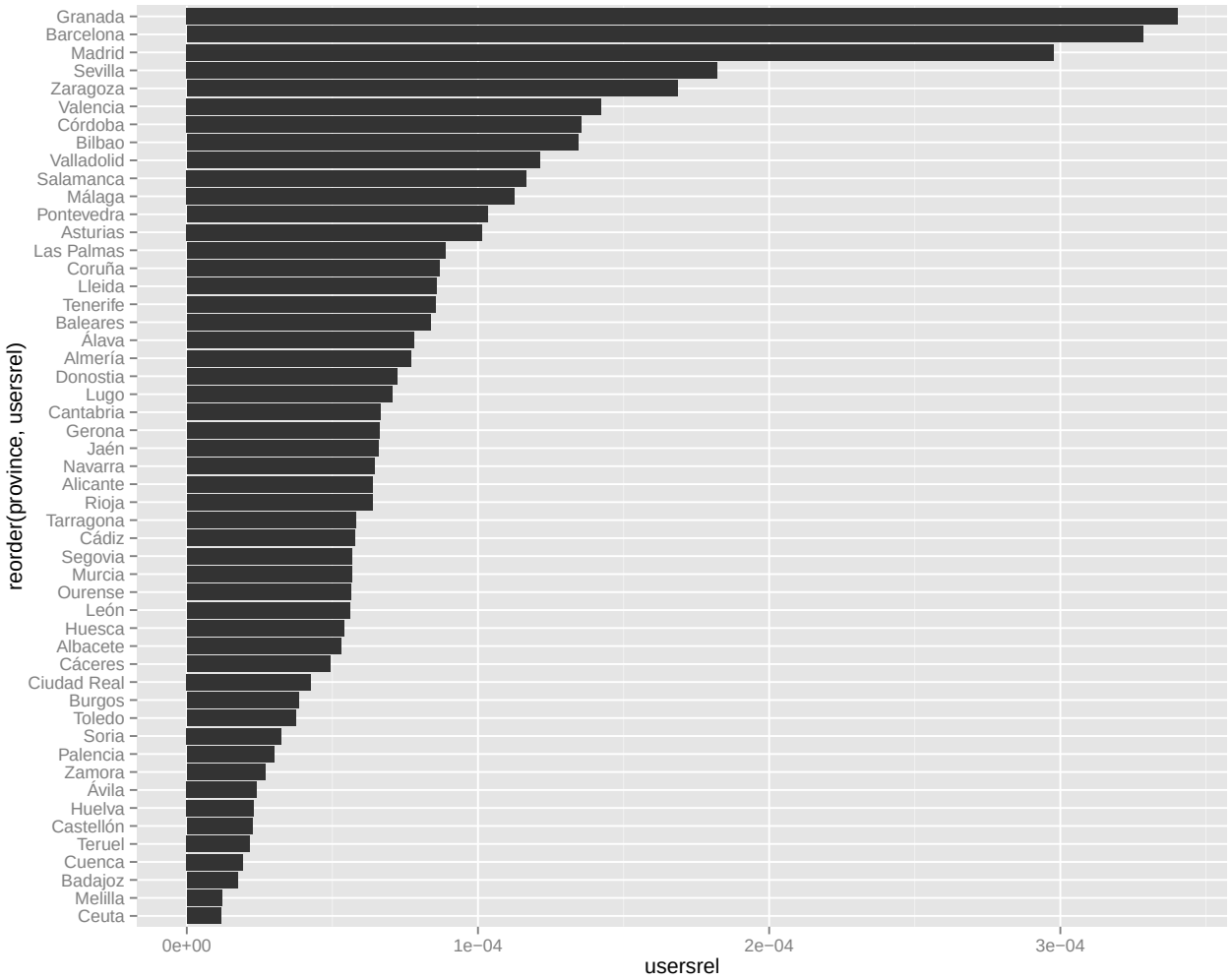


This graph does not include the non-provincial users, that, is, those whose province was not declared; besides, it does not give you an idea of the amount of users per province *relative* to the total number of users, which is around 10000. This is shown next

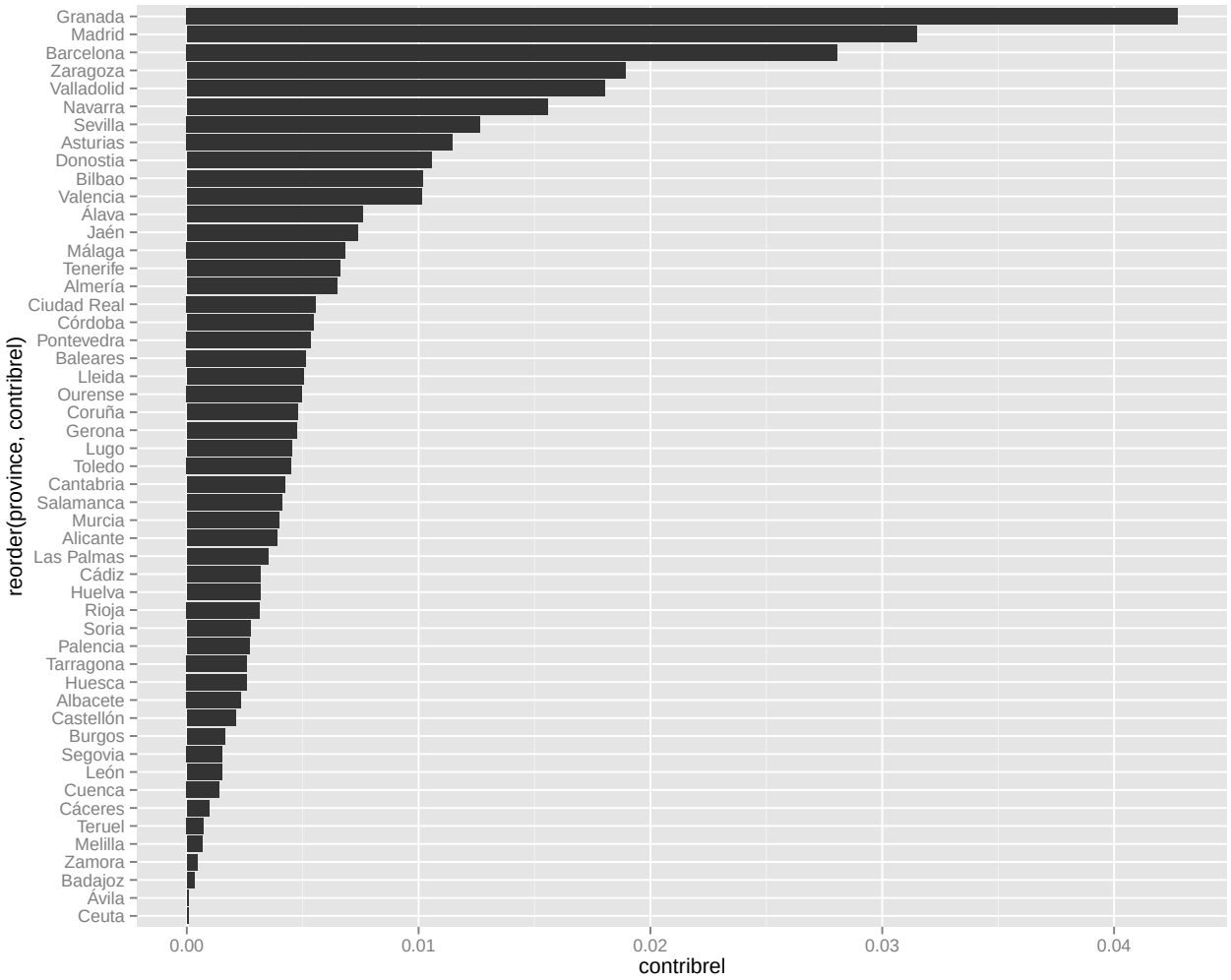


The users without a declared province form the biggest slice of the pie, with around one quarter of the total number of users. They are followed by Madrid, Barcelona, Valencia, Seville, Málaga and Granada. These provinces, by themselves, host more than half the total community of GitHub users in Spain.

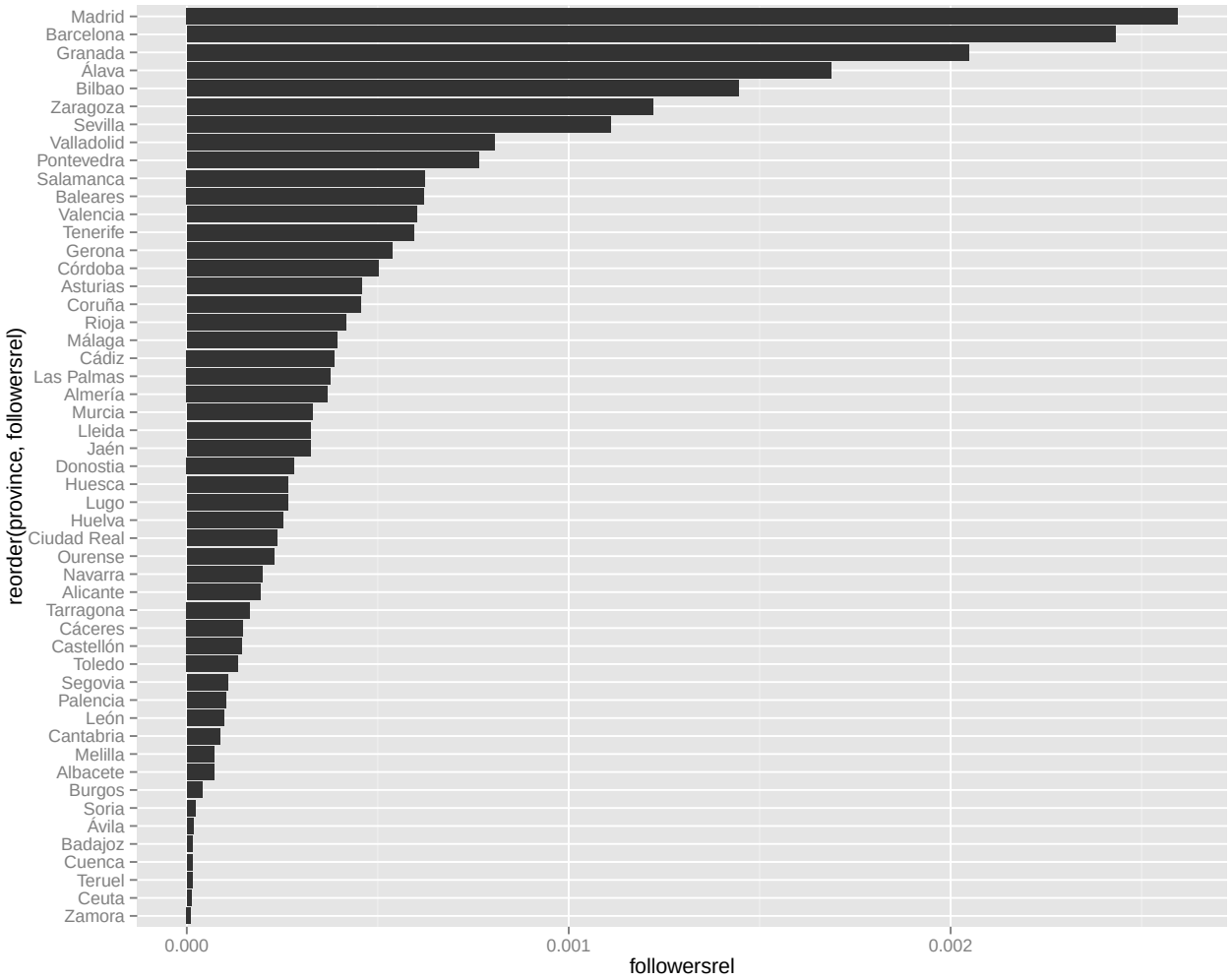
Except for Valencia, Barcelona and Madrid, the rest of the provinces are not the most populated in Spain. That is why, if we take into account the population, dividing the number of GitHub users by the provincial population (as published by the National Statistics Institute), the situation is somewhat different, with Granada emerging as the province with the highest number of GitHub active developers per capita, followed by Barcelona and Madrid and then by Seville and Saragossa.



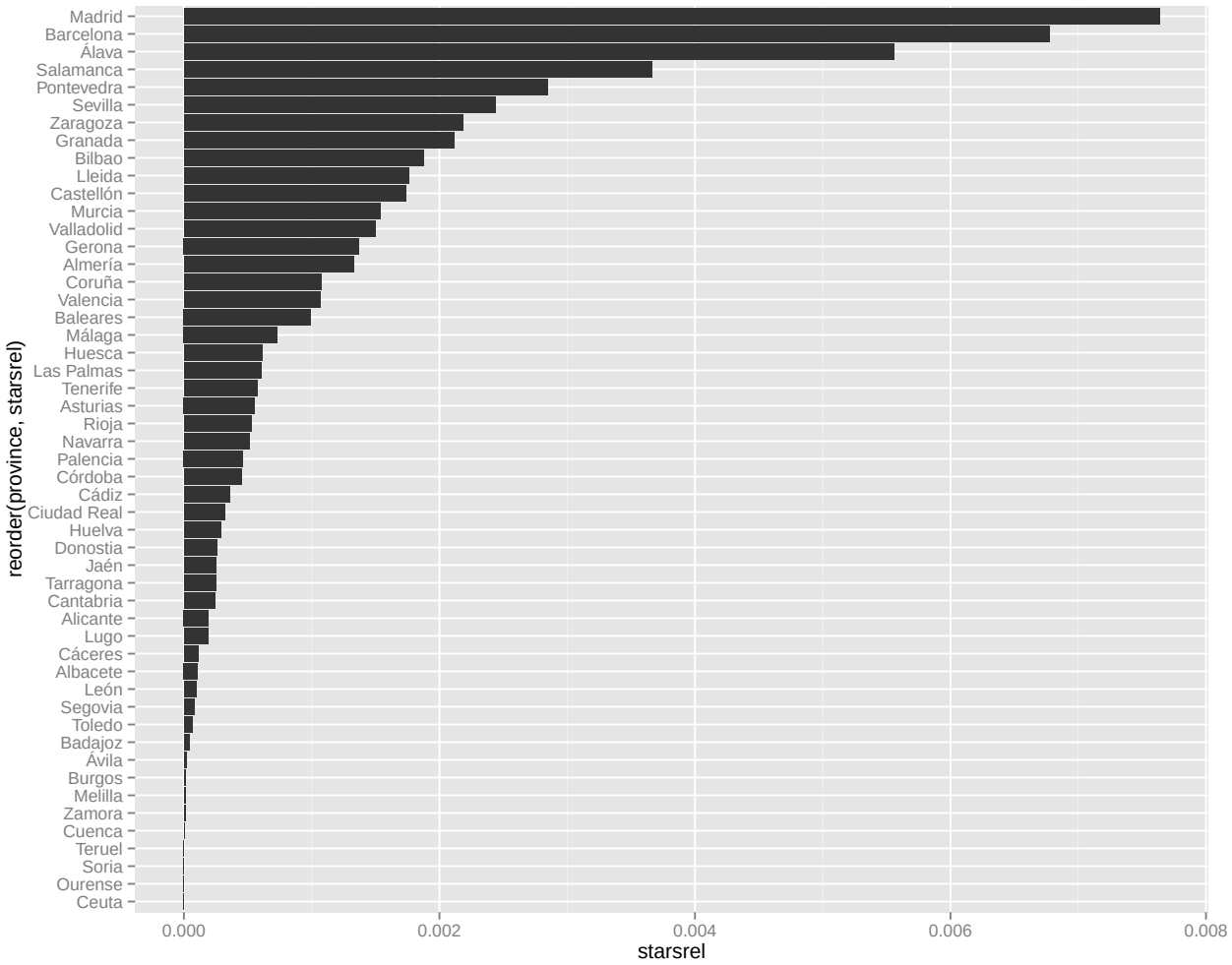
The situation is similar if we take into account the aggregated contributions by all province users. Once again, Granada emerges as the winner but the position of Barcelona and Madrid is inverted and a new player, Valladolid, enters the top five.



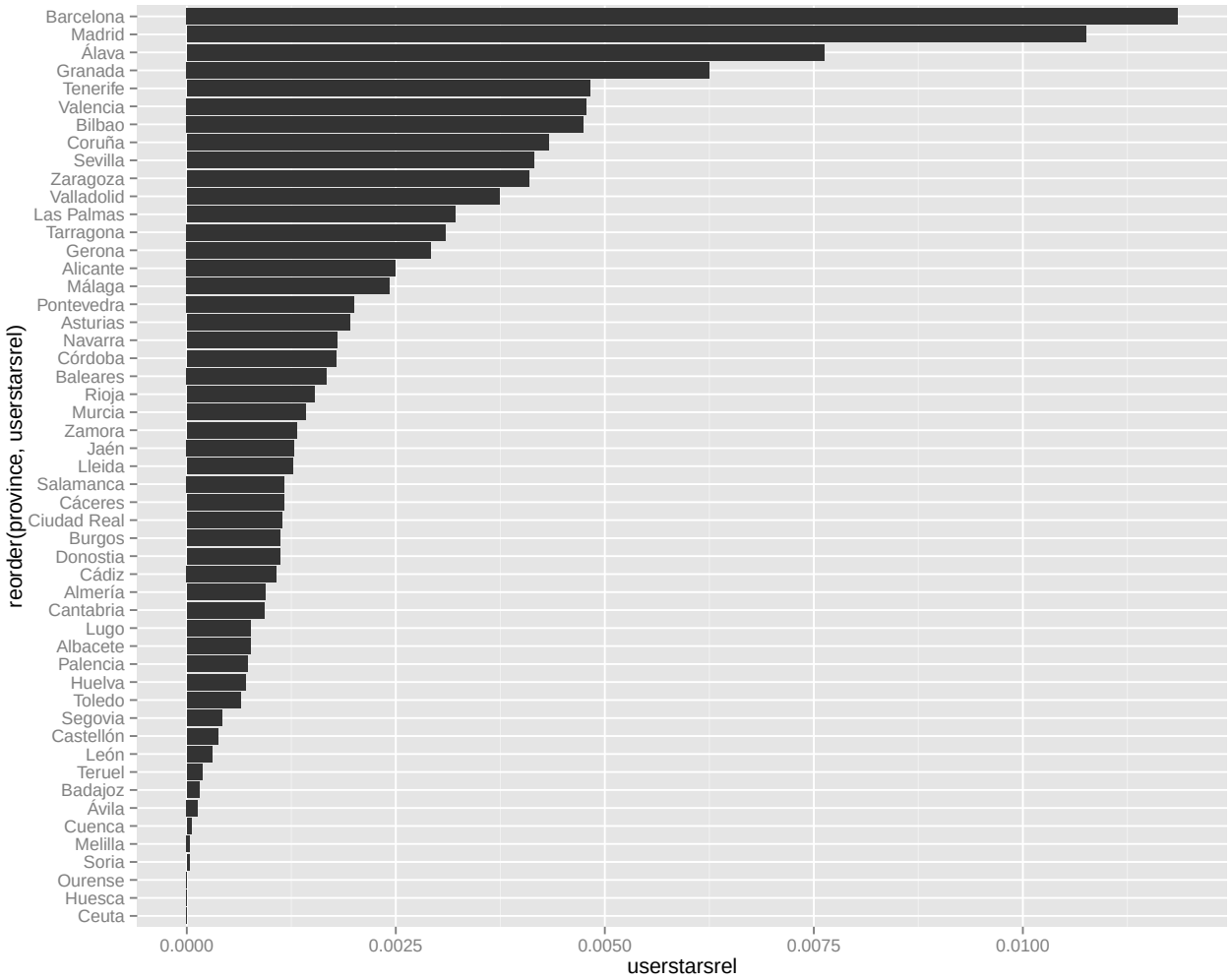
The aggregated number of followers is a bit less surprising, with Madrid and Barcelona on top, but Álava and Bilbao entering the top 5. If we delve into data this is mainly due to a single user in each case.



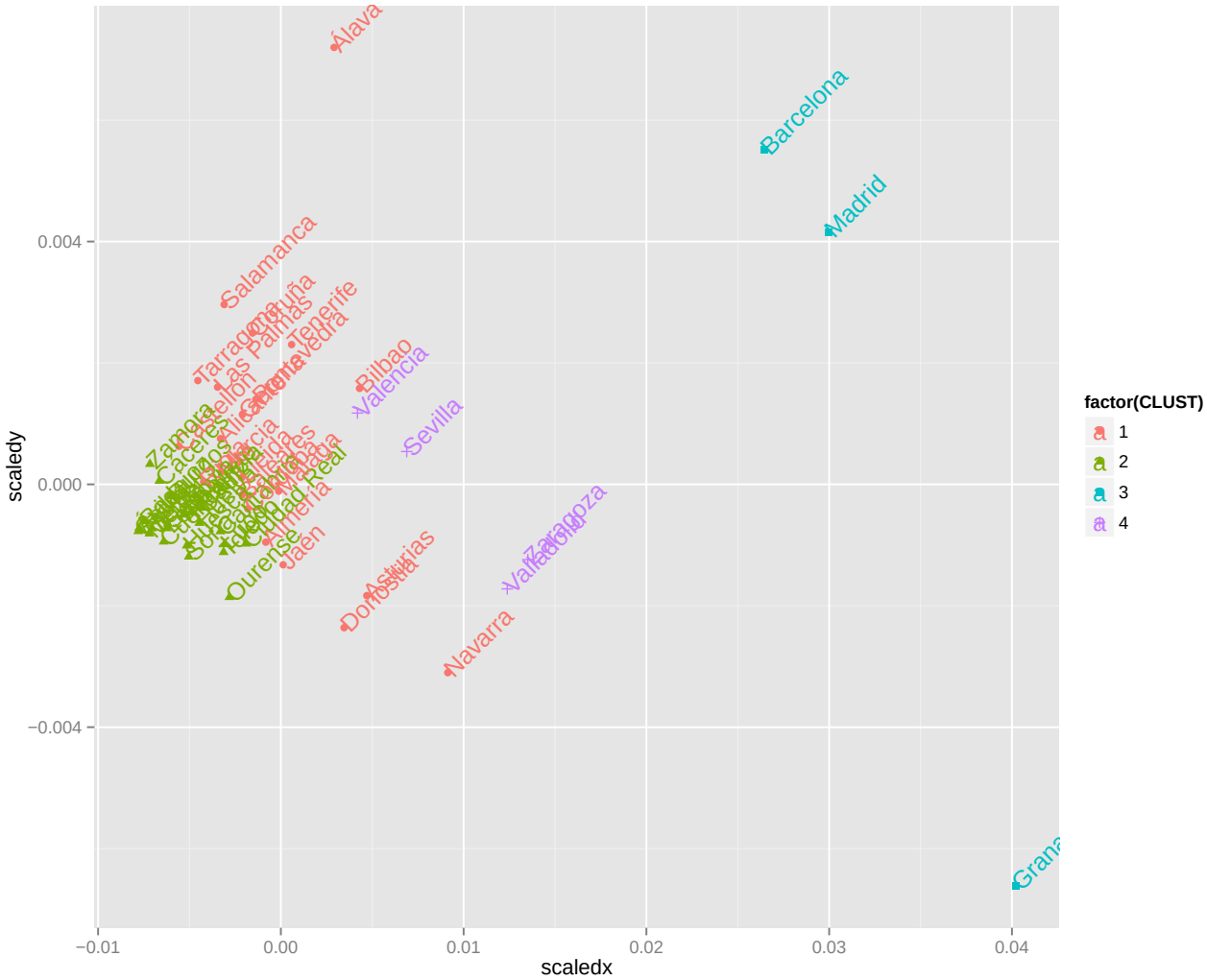
The stars given to projects, which is a proxy for popularity, is correlated (but it remains to be seen exactly how) to the number of followers, with Álava again on the top 5 and two completely new provinces, Salamanca and Pontevedra, getting to the top.



Finally, if we consider the number of stars given by users, a new one, Tenerife, gets into the top 5.



The graphs above imply that there are differen classes in the provinces in Spain. We have performed clustering using mclust (Fraley and Raftery (2002),Fraley et al. (2012)) using as representation for each province the relative values plotted above and obtained this division into four clusters.



Apparently, one cluster includes the provinces with the highest productivity: Madrid, Barcelona and Granada. A second cluster, in purple, includes Valencia, Sevilla, Zaragoza and Valladolid, which consistently score in the top 10 in every ranking and in some cases the top 5. The following cluster includes many provinces like Álava or Navarra which show up in some cases in the top 5 for some reason, but not always. The rest of the provinces are grouped in the remaining cluster, shown in green.

Conclusions

This paper measures and ranks Spanish provinces by the number of users and other quantities related to productivity (contributions) and popularity (stars, followers). It is a first approximation to community metrics in Spain and it is mainly intended as a reference for future use. It is also an indication of a particular point in time. Future versions will probably change this scenario and it is an interesting line of work to delve on the reason for these changes.

Acknowledgements

I am grateful to Francisco Charte for his help creating the pie chart for this paper.

References

Fraley, Chris, and Adrian E. Raftery. 2002. "Model-Based Clustering, Discriminant Analysis and Density Estimation." *Journal of the American Statistical Association* 97: 611–31.

Fraley, Chris, Adrian E. Raftery, Thomas Brendan Murphy, and Luca Scrucca. 2012. *Mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation.* Technical Report. Department of Statistics, University of Washington.

Merelo, J. J. 2015. *GitHub Users in Spain: An Initial Analysis.* GeNeura Team http://geneura.wordpress.com; RPubs, http://rpubs.com/jjmerelo/gh-users-spain.

Merelo-Guervos, Juan-Julian, Israel Blancas, Maribel G. Arenas, Fernando Tricas, José Antonio Vacas, and Nuria Rico. 2015. "GitHub Rankings and Its Impact on the Local Free Software Development Community." *The Winnower*, January. doi:10.15200/winn.142251.14740.