Title:
The role of undocumented cases in the propagation of the COVID-19 virus in Spain

Authors and affiliations and contact emails
Luis Orea, Universidad de Oviedo, lorea@uniovi.es
Inmaculada C. Álvarez, Universidad Autónoma de Madrid, inmaculada.alvarez@uam.es

Abstract
This research extends a previous article, published in FEDEA (http://documentos.fedea.net/pubs/dt/2020/dt2020-03.pdf) where we examined the effectiveness of the Spanish lockdown of the 14th of March. This study ignored the existence of undocumented coronavirus cases. This is an important issue that should be addressed, first because the proportion of COVID-19 infections not detected by the Spanish health systems is much larger than the proportion of laboratory-confirmed coronavirus cases (see Flaxman et al. 2020), and second because the undocumented cases facilitate the rapid dissemination of this new coronavirus (see Li, 2020). We here consider the cumulative number of both laboratory-confirmed and undocumented coronavirus cases. We use the stochastic frontier analysis approach to control for undocumented cases because these infections are not observed by the econometrician and the reported cases are always lower than the total number of COVID-19 infections. In this sense, the unobserved cases can be proxied using a one-sided random term in the same fashion as firms’ inefficiency in production economics. This is the main methodological contribution of this paper. Another contribution has to do with the spatial specification of our model. Following our previous paper and the scarce epidemiology literature that controls for spatial propagation effects (a remarkable exception is Giuliani et al., 2020), we next assume that the total number of infected persons in province $i$ depends not only on laboratory-confirmed cases but also on undocumented coronavirus cases in neighboring provinces. In this sense, we define a model that captures inter province connections in terms of geographical proximity, affective links, student’s mobility and tourist habits to capture the propagation of the virus between provinces. Estimating this model presents, however, important methodological challenges due to the spatial correlation of the one-sided random term capturing the unobserved cases.

Data description
We create a daily data base of laboratory-confirmed COVID-19 cases in the Spanish provinces covering the period from the onset of the epidemic in each province to the 4th of April, constructed from several sources. This data was collected manually by the authors from the official press releases of the Spanish regional governments, the Ministry of Health and Wikipedia. In particular, we had to consult these information sources to extend backwards the provincial data published by Datadista in GitHub under a free License since 13th March,14 which extracts their data from a variety of documents published by the Ministry of Health. From the 28th March onwards we collected the data directly using RTVE Flourish, a tool that creates high-end maps and summaries the information of each province. We used the regional online data released by the Ministry of Health to check the information provided by Datadista and RTVE Flourish. We have also used the region-level data released by the Ministry of Health and the province-level dada released by the Spanish regional governments to correct typos and lack of information on coronavirus cases in some provinces (e.g. in Galicia).
We consider several sources of spatial links to analyze the expansion of the virus. Therefore, we define alternatives $W$ weighted matrices: affective links with city residents, mobility of students, transport connections and tourist habits. These matrices are provided by different sources, such as INE, Ministry of Education, Ministry of Transport and the Ministry of Industry, Trade and Tourism.

JEL Codes for the project
I1, H840, Q54, R12, C23

Key-words
COVID-19, Spanish lockdown, spatial propagation, Stochastic Frontier Analysis