
# EXTENDED ABSTRACT

## Estimating a stochastic epidemic frontier model that controls for undocumented COVID-19 cases with an application to Spain

**Luis Orea**
(lorea@uniovi.es) Departamento de Economía, Universidad de Oviedo
**Inmaculada C. Álvarez**
(inmaculada.alvarez@uam.es), Departamento de Análisis Económico: Teoría Económica e Historia Económica, Universidad Autónoma de Madrid
**Alan Wall**
(awall@uniovi.es), Departamento de Economía, Universidad de Oviedo

**Subject area:** 6. Methods for regional and urban analysis: big data, spatial econometrics and geographic information systems

**Abstract:** *(minimum1500 words)*

The COVID-19 pandemic, which began in China in December 2019, spread worldwide in a short time. Faced with the threat of their public health systems being overwhelmed, several countries, with Italy and Spain at the forefront as they were the most-affected at the initial stage of the pandemic, saw themselves forced to implement national lockdowns of the population.

The effectiveness of such institutional responses to contain the propagation of this disease has given rise to a rapidly-evolving literature. One of the first studies to examine the effectiveness of the control measures implemented in several European countries was carried out by Flaxman et al. (2020). They find that the Spanish lockdown averted about 67% of potential deaths by the 31st of March, 2020. Using data from 34 European countries, Cho (2020) has recently found that the infection cases in Sweden would have been reduced by almost 75% had its policymakers followed stricter containment policies. Examinations of the effectiveness of institutional control measures while controlling for spatial propagation effects has been treated only marginally in the literature. A notable exception is Dickson et al. (2020) who find that in the northern Italian provinces the Government containment measures not only succeeded in drastically reducing the transmission of COVID-19 amongst individuals within these provinces, but also avoided contagions between neighbouring areas. Another exception is Orea and Álvarez (2020), who find similar results using a simple but novel empirical strategy to capture the typical *S*-shaped temporal pattern of the virus epidemic when examining the propagation of COVID-19 across the Spanish provinces.


Aside from spatial propagation effects, another important issue that has often been overlooked in this literature is the number of undocumented coronavirus cases. The relevance of this lies in the fact that the proportion of coronavirus infections not detected by the health system during the first wave of contagion of COVID-19 was likely much larger than the proportion of laboratory-confirmed coronavirus cases. As Korolev (2021) points out, if we do not take underreporting into account and estimate models from data on confirmed cases under the assumption that all cases are reported, our estimates might be seriously biased.

With this in mind, in this paper we estimate a stochastic epidemic frontier (SEF) model in order to account for the prevalence of undocumented cases, which are not observed by the econometrician. As the reported cases are always lower than the total number of infections, the unobserved cases can be proxied using a one-sided random term in the same fashion as firms' inefficiency in the stochastic frontier analysis (SFA) literature.[1] Our SEF model permits reporting rates to be estimated rather than assumed (see e.g. Chudik et al., 2020) and is flexible enough to permit these reporting rates to vary across geographical cross-section units of observation. Our work also provides a methodological contribution for practitioners aiming to estimate firms' efficiency in a production economics setting because it provides a heteroskedastic version of the model introduced by Wang and Ho (2010).

We provide an empirical application of our models to Spanish data corresponding to the initial months of the original outbreak of the virus in early 2019. We have used several sources to construct a dataset of coronavirus cases across Spain. As most control measures began on the days of March 13th and 14th, 2020, we analyse data on coronavirus cases two weeks before and two weeks after those dates. In particular, our data set covers the period between the onset of the epidemic in each province and the 4th of April.

In all estimated models, we find very different reporting rates across the Spanish provinces. The large cross-sectional heterogeneity in reporting rates found in our empirical application is one of the contributions of the paper, as previous epidemiological literature has often assumed common rates. On average, most of our reporting rates range from 10% to 79%, a similar variation found for the exposure rates in Chudik et al. (2020). Therefore, it may be the case that their estimated variety of exposure rates is caused by of the fact that their econometric model ignores systematic variations in reporting rates across provinces.

Most of our estimated models provide evidence that the exodus from the epicentres of the Spanish coronavirus crisis of people wishing to spend the lockdown in provinces with few or no cases of COVID-19 markedly spread the virus across the country. Therefore, restricting people's mobility (between or within provinces) seems to be a reasonable measure to attenuate the propagation of the coronavirus. In this sense, our results show that the lockdown was effective both in preventing the propagation of the coronavirus *between* provinces as well as in attenuating the propagation of the virus

[1] In current unpublished work, Millimet and Parmeter (2020) also propose a stochastic epidemic frontier model. Our approaches are quite different, however. For example, their SFA model focuses on *new* coronavirus cases, whereas we focus our model on *cumulative* cases. While they also examine measurement issues with the number of deaths, their model does not have an autoregressive structure and does not account for spatial propagation.

*within* each province. Hence, we find that the Spanish lockdown, together with other control measures, was an effective measure to battle COVID-19 in the absence of pharmaceutical measures (e.g., vaccines).

The average contraction in the rates of growth of coronavirus cases attributed to the lockdown is around 6.8 percentage points (from 18.2% with no lockdown to 11.4% with the lockdown). The largest reductions were found in provinces that are either close to the epicentres of the coronavirus or adjacent to provinces with more advanced epidemics. The reductions in the rates of growth of coronavirus cases attributed to the lockdown in these provinces are much larger than the average value. For instance, we find notable effects in Ávila, Segovia and Cuenca, which neighbour Madrid, the Spanish province hardest-hit by coronavirus. Large effects are also found in Tarragona and Lérida, which neighbour Barcelona, the second hardest-hit Spanish province. We also find large effects of the lockdown in Ciudad Real and Albacete, two adjacent provinces that are two local foci of the coronavirus in the centre of Spain. In southern Spain, we find large effects in Córdoba, which neighbours Málaga, the main epicentre of the coronavirus in this area. We also find important effects for sparsely-populated provinces such as León, Soria, Palencia, Burgos and Teruel. It is worth mentioning that the epidemic in many of these provinces began almost one week later than it did in neighbouring provinces. Therefore, while local and national lockdowns of the population are effective measures to battle COVID-19, they should be implemented at the very early stages of the epidemics.

We also extended our pure frontier epidemic models by including a set of socio-economic factors that might influence the evolution of the epidemic in each province. This information can be very useful for policy makers and health authorities planning the relaxation of a lockdown. We find that the most-populated provinces had more intensive coronavirus epidemics. More (less) intensive coronavirus epidemics were also found in provinces with a relatively large share of workers in the service (agricultural) sector. These results, together with the strong propagation effects estimated for provinces close to the main epicentre of the coronavirus in Spain, poimt to the idoneity of carrying out a gradual, focused relaxation of the control measures. Thus, the relaxation of the lockdown should be slow in the most-populated provinces, in provinces with a higher share of the workforce in the service sector, and in the main epicentres of the coronavirus of Spain. Control measures could be lifted earlier in provinces mainly engaged in primary-sector production. Finally, our findings also suggest prioritizing the detection of coronavirus cases at early stages of the epidemics as an effective strategy to combat the propagation of this virus.

The models presented permit undocumented cases to be estimated, rather than assumed, and also allow spatial propagation of the virus across geographical areas to be modelled. A simulation exercise indicated that the epidemic-time model performed better, and in an empirical application to the case of the original outbreak of the pandemic in Spain we provide estimates from several different specifications of this model. The results from our models provided insights into the effectiveness of the national and regional lockdown measures and the influence of socio-economic factors in the propagation of the virus.

Our work can be extended in several directions. In the empirical application in this paper we availed of data at provincial level that allowed us to analyse the effectiveness of national and regional institutional responses at this level of disaggregation. However, several regions in Spain, including Andalusia, Asturias, the Basque Country, Cantabria, Catalonia, Madrid and Murcia have also provided data on coronavirus cases at municipal level. By adapting our empirical strategy to this more disaggregated data we will be able to evaluate the local control measures established by the regional governments during the second and successive waves of contagion of COVID-19.

Another extension would be to explore the possibility of different collectives within the population having different proportions of asymptomatic or undocumented cases. For example, data at provincial level by gender would allow us to examine whether the proportion of undocumented cases among women is larger or smaller than that among men. If this were the case, public health authorities should be particularly aware of gender-based channels of transmission of the virus in sectors of the economy where one gender or the other makes up a substantial majority of the workforce. These types of differences between collectives can be modelled with a system of epidemic spatial stochastic frontier equations, one for each collective. The copula-based maximum likelihood (ML) approach introduced by Lai and Huang (2013) is well-suited for such an analysis.

Finally, the relationship between reported and undocumented cases could be explored in greater depth. Li et al (2020) have indicated that undocumented (asymptomatic) cases facilitate the dissemination of COVID-19. It is not clear how to explore this cross-group propagation effect using a frontier analysis approach because it tends to “reverse” the sign of the one-sided error term capturing the proportion of undocumented cases. A candidate is the latent class frontier model approach of Kumbhakar et al. (2007), as this model allows the sample to be split into two groups that differ in how the one-sided error term enters the model.

**Keywords:** SIR models, stochastic frontier analysis, panel data, COVID-19, Spain.
**JEL codes:** C23, C51, I18