ECONOMIC DISCUSSION PAPERS

Efficiency Series Paper 1/2021

Estimating the propagation of the COVID-19 virus with a stochastic frontier approximation of epidemiological models: a panel data econometric model with an application to Spain

Luis Orea, Inmaculada C. Álvarez, Alan Wall

Departamento de Economía

Universidad de Oviedo

Available online at: https://www.unioviedo.es/oeg/

Estimating the propagation of the COVID-19 virus with a stochastic frontier approximation of epidemiological models: a panel data econometric model with an application to Spain

Luis Orea *

Department of Economics, University of Oviedo and Oviedo Efficiency Group

Inmaculada C. Álvarez

Department of Economics, Universidad Autónoma de Madrid and Oviedo Efficiency Group

Alan Wall

Department of Economics, University of Oviedo and Oviedo Efficiency Group

January 24, 2021

Abstract

The literature examining the propagation of COVID-19 has mainly used pure epidemiological models focused on estimating reproductive numbers, mortality and other epidemiological features. In this paper we use a stochastic frontier analysis (SFA) approach to model the propagation of the epidemic across geographical areas, which complements existing epidemiological models. Our work bridges the SFA and epidemiological literatures and shows that the translation from epidemiological models to SFA implies strong assumptions and introduces measurement errors. We propose two different specifications of the stochastic frontier model: first, a stochastic frontier based on an epidemiological SIR model specification; and second, an approximation to this SIR-based frontier based on functions of the length of time since the outbreak of the virus began. These models permit reported and undocumented cases to be estimated. The appeal of these models lies in the fact that they can be estimated using only epidemic-type data and yet are flexible enough to permit these reporting rates to vary across geographical cross-section units of observation and to allow other covariates affecting reported and undocumented rates to be incorporated. 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 where we introduce a series of series of extensions to base model and specification robustness checks.

Keywords: SIR models, stochastic frontier analysis, panel data, COVID-19, Spain.

JEL: C23, C51, I18

---

* *Corresponding author at: Department of Economics, University of Oviedo, School of Business and Economics, Avda. del Cristo s/n, 33006, Oviedo, Spain. Phone: +34 985106243. Email addresses: lorea@uniovi.es.*

1. Introduction

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. In the specific case of Spain, this gave rise to heated debates, which would be repeated in other countries (notably the UK), over the timing and duration of the lockdown. There was fierce criticism from some opposition parties over the Spanish national government's handling of the first wave of the pandemic, and it is noteworthy that the institutional response in Spain to the second wave which began during the autumn of 2020 has been delegated to regional governments which are charged with implementing measures at local or regional level. A consequence of the regional nature of the new institutional response, however, is that much less attention may be paid to the propagation of the coronavirus across the Spanish provinces and regions.

The propagation of the COVID-19 epidemic and the effectiveness of institutional responses has given rise to a rapidly-evolving literature. Most of this literature (see Millimet and Parmeter, 2020 for a survey) has focused on estimating reproductive numbers, mortality and other epidemiological features. One of the first studies that aimed 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. Regarding the Chinese COVID-19 epidemic, Leung et al. (2020) find that a relaxation of the control measures in force in China would increase the cumulative number of coronavirus cases, bringing forward a possible second wave. These authors conclude that it is necessary to monitor the increase in new cases due to the effects of relaxing control measures in order for policy makers to be able readjust their decisions.

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 Gross et al. (2020), who study the spatio-temporal propagation of COVID-19 in China and compares it to other countries. They conclude that early action may attenuate the disease, given the strong relation between population migration and the spreading of disease. Giuliani et al. (2020) also use data disaggregated by provinces to implement an epidemiological model explaining the propagation of COVID-19 across the Italian provinces. The origin of this spatial dimension of propagation is the high inter-provincial mobility of people. They conclude that the control measures were more successful in those provinces with more effective enforcement.

Aside from spatial propagation effects, another important issue that has often been overlooked or not controlled for 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 (see Flaxman et al., 2020), with the result that the official number of coronavirus cases likely falls short of the true number of cases, perhaps significantly so. 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. In addition, underreporting may dampen public and political support for more stringent measures such as investments in medical equipment, mandatory masks or mandatory lockdowns.

To account simultaneously for geographical propagation of the virus, the prevalence of undocumented cases and the effectiveness of institutional control measures, in this paper we propose a stochastic frontier analysis (SFA) approach to estimating epidemic curves. The SFA approach can be used to control for the existence of undocumented coronavirus cases because these cases are not observed by the econometrician and the reported cases are always lower than the total number of COVID-19 infections. Therefore, the unobserved cases can be proxied using a one-sided random term in the same fashion as firms' inefficiency in production economics. The model we propose can be seen as an extension to a frontier setting of previous work by Orea and Álvarez (2020), which examined the propagation of COVID-19 across the Spanish provinces and assessed the effectiveness of the first Spanish population lockdown of population. Although they only estimated the epidemic curve of reported COVID-19 cases, Orea and Álvarez (2020) introduced a simple but novel empirical strategy to capture the typical *S*-shaped temporal pattern of the virus epidemic.

The added value of our paper is combining the stochastic frontier analysis and the classical Susceptible-Infected-Removed (SIR) epidemiological model. In our attempt to bridge the epidemiological modelling and stochastic frontier literatures, we provide two different specifications of the epidemic stochastic frontier analysis model (ESFA). The first is a stochastic frontier inspired by the non-frontier econometric SIR model proposed by Chudik et al. (2020), which we denote as the *SIR-based model*. The second is an approximation to this SIR-based frontier that replaces the time-varying epidemiological regressors with functions of the length of time since the outbreak of the virus began, an approach used by Orea and Álvarez (2020) to capture the shape of the epidemic curves. We label this the *epidemic-time model*.

The simple model specifications we propose have a number of appealing features. The first of these is that the models rely on relatively little information, in that both the epidemic-time and SIR-based specifications of the stochastic frontier model can be estimated using epidemic-type data only, i.e., the rates of growth of coronavirus cases depend in our models on own and neighbours' epidemic times, lagged cases of COVID-19, date of implementation of control measures, and so on. However, the model is flexible enough to include other covariates if deemed appropriate. Another advantage of our model is that it permits reporting rates to be estimated rather than assumed and is flexible enough to permit these reporting rates to vary across geographical cross-section units of observation. As such, our ESFA models can be thought of as complementary to existing epidemiological models, such as Chudik et al. (2020), which often assume common reporting rates across areas.[1]

As the volatility of the rates of growth of reported cases are typically much larger at the beginning of the epidemic than when the epidemic has advanced, our ESFA model must be estimated using time-varying heteroskedastic noise terms. To capture this feature, we propose a stochastic frontier specification which can be interpreted as a heteroskedastic version of the model introduced by Wang and Ho (2010) whose aim was to control for individual effects in a production economics setting. Therefore, our paper has also a methodological contribution for practitioners aiming to estimate firms' efficiency using the Wang and Ho (2010) approach.

[1] For example, Chudik et al. (2020) use the data from the Diamond Princess cruise ship reported by Moriarty et al. (2020) to calibrate the proportion of the population exposed to COVID-19, and assume an average reporting rate in all Chines provinces of 50%. They find large variations in exposure rates across Chinese provinces, ranging from 9% to 87%. The fact that their econometric model ignores systematic variations in reporting rates across provinces may well be causing this wide variety of exposure rates.

Estimating an epidemic stochastic frontier analysis model mimicking *all* features of a SIR model presents important methodological challenges because the standard stochastic frontier estimators in production economics do not capture the complexities of the SIR model. When trying to specify a tractable ESFA model, we find that some simplifying assumptions need to be made. To test the implications of this, we carry out a simulation analysis where we check the performance of the epidemic-time and SIR-based specifications of the stochastic frontier model, finding that the epidemic-time specification performs better. In our empirical application using panel data from the Spanish provinces observed over the initial period of the COVID-19 outbreak in Spain, we corroborate that the epidemic-time model provides more realistic estimates of reporting rates than the SIR-based frontier specification.

As our epidemic-time model can be extended to include other covariates, in our empirical application we take advantage of this feature to incorporate a series of socio-economic and environmental variables to test their influence of the evolution of total and under-reported cases. We also carry out a series of robustness checks on our epidemic-time stochastic frontier model, including an analysis of the effects of changes to the distributional assumptions and the effects of changing the actual panel data set used to check the effect of dropping observations with zeroes in the variables.

Overall, the empirical strategy used in this paper can be said to rely several different but related assumptions, which are supported by previous literature: i) the propagation of the virus across areas (Spanish provinces in our application) depends on people's mobility (Giuliani et al., 2020); ii) this mobility can be modelled using spatial econometrics techniques (Eliasson et al., 2003; Orea and Álvarez, 2020); iii) the undocumented cases represent a large proportion of total cases of infection (Flaxman et al., 2020); iv) the proportion of undocumented cases through the epidemic development varies over time (Li et al., 2020); and v) the unobserved cases can be proxied using a one-sided random term (Millimet and Parmeter, 2020).[2]

The paper proceeds as follows. Section 2 defines the three epidemic curves we use, namely the total epidemic curve, the reported cases epidemic curve, and the undocumented cases epidemic curve. In Section 3 we present the stochastic frontier representation of the epidemic curves. Distributional assumptions about the error terms and the maximum likelihood procedure for the general specification of the model are discussed. This section concludes with a presentation of alternative specific specifications of the frontier model, namely a SIR-based frontier model and its epidemic-time approximation. The performances of these two specifications are compared through a simulation exercise. Section 4 presents our empirical application to Spanish provinces at the outset of the COVID-19 epidemic in the spring of 2020. We first estimate a basic version of our preferred frontier model, namely the epidemic-time model, with a spatial lag specification. We compare our results to those from a similar SIR-based specification and then discuss a series of extensions and robustness checks. Section 5 concludes.

[2] In current unpublished work, Millimet and Parmeter (2020) propose a stochastic frontier model also based on the classical SIR 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.

2. Total and partial epidemic curves.

In this section we define three epidemic curves that resemble the popular reproduction-based models used in the epidemiological literature, which often ignore the existence of undocumented coronavirus cases.

Consider a panel of $i = 1, \dots, N$ provinces observed on $t = 1, \dots, T$ days. t is the calendar time. Let E_i denote the *onset date* of the epidemic, namely the date on which province i reports its first coronavirus case. We then analyse the development of the epidemic in each province, i.e., the temporal evolution of coronavirus cases once each province reports its first coronavirus case. A key variable to carry out this analysis is the *epidemic time*, $K_{it} = t - E_i$, which denotes the number of days since the onset date. Next, let Y_{it}^* denote the *accumulated* number of both laboratory-confirmed (Y_{it}) and undocumented (U_{it}) coronavirus cases until day t in province i. Thus:

$$Y_{it}^* = Y_{it} + U_{it} \tag{1}$$

In Orea and Álvarez (2020), the epidemic curve of reported cases (Y_{it}) is represented by an autoregressive relationship:[3]

$$Y_{it} = \beta_{it} Y_{it-1} \tag{2}$$

where β_{it} can be interpreted as an autoregressive parameter (function) that depends on a set of covariates. We label this the *epidemic curve*. The key aim of coronavirus control measures is to reduce β_{it}. If β_{it} is close to 1, the number of new infections is relatively small and the epidemic has therefore been controlled. If β_{it} is much greater than 1, then a lot of new infections have been reported and the coronavirus epidemic is still spreading among the population despite the efforts to prevent the propagation of the virus. The β_{it} parameter (function) thus plays the same role as the so-called *reproductive number of the infection* (R_0), a fundamental quantity used in the epidemiological literature to represent the average number of infections per infected person over the course of their infection.

In order to get a simple empirical specification of (2), we take natural logarithms and first-differentiate the model.[4] This yields the following expression:

$$\Delta ln Y_{it} = ln Y_{it} - ln Y_{it-1} = ln \beta_{it} \tag{3}$$

where $ln\beta_{it}$ simply measures the daily rate of growth of reported cases. We expect rates of growth of coronavirus cases to vary with the epidemic time, K_{it}, because the traditional epidemic curve has a S-shaped form. If this is indeed the case, the epidemic curve β_{it} can be modelled empirically as a third-order function of the (logged) epidemic time variable, conditional on other control variables.[5]

[3] The model that describes the expected number of infections at time (day) t in Giuliani et al. (2020) is also allowed to depend on the number of infections reported at time $t - 1$.

[4] We have found in our application that Y_{it} is not a stationary variable. Estimating (2) might thus give spurious results. This issue vanishes if we use rates of growth of reported coronavirus cases.

[5] Figure 3, which shows the box plots of the rates of growth of reported cases by epidemic time, clearly reveals that the rates of growth of reported cases are much larger at the beginning of the epidemic than when the epidemic

Similar autoregressive expressions can be written for undocumented and total coronavirus cases. That is, each variable measuring coronavirus cases has its own epidemic curve. While the epidemic of reported cases is given by (3), the epidemic curves of undocumented and total coronavirus cases can be written as follows:

$$U_{it} = \theta_{it} U_{it-1} \tag{4}$$

$$Y^*_{it} = \beta^*_{it} Y^*_{it-1} \tag{5}$$

Figure 1 illustrates our three hypothetical epidemic curves. By construction, we have assumed in this figure that Y^*_{it} is the sum of Y_{it} and U_{it} for each epidemic time K_{it}. Note that while the epidemic curve of reported cases has the traditional S-shaped form, the epidemic curve of undocumented cases is depicted using a log form from the beginning of the epidemic onwards. This allows the proportion of undocumented cases to decrease over time as in Li et al. (2020). The shape of the total epidemic curve is thus a combination of the shapes of the two partial epidemic curves.

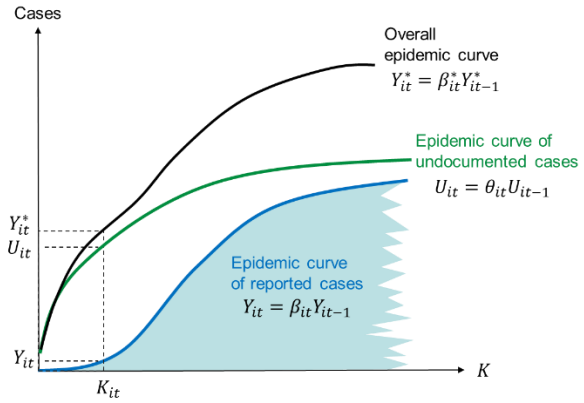

Figure 1: Epidemic curve of total, reported and undocumented cases

We now examine this feature analytically. Taking into account (1), the autoregressive parameter β^*_{it} can be decomposed as follows:

$$\beta^*_{it} = \beta_{it} + (\theta_{it} - \beta_{it}) U_{it-1} / Y^*_{it-1} \tag{6}$$

This equation shows that the overall epidemic curve coincides with the epidemic curve of reported cases if both reported and undocumented cases have the same temporal patterns (i.e., $\theta_{it} = \beta_{it}$). In order to link both epidemic curves, let u_{it} denote the log difference between total and reported coronavirus cases:

$$u_{it} = lnY^*_{it} - lnY_{it} \tag{7}$$

Given the above definition, the proportion of undocumented cases can be expressed as a increasing function of u_{it} because $U_{it}/Y^*_{it} = 1 - e^{-u_{it}}$. u_{it} can therefore be viewed as a relative measure of the undocumented cases in an epidemic outbreak: loosely speaking, we can

has advanced. That is, $ln\beta_{it}$ tends to decrease rapidly in the early stages of the epidemic. This figure also shows a flattening in mid-stages of the epidemic and very small rates of growth in the later stages of the epidemic.

interpret u_{it} as the "proportion of undocumented cases". Equation (7) also allows us to link the reported and undocumented cases as follows:

$$U_{it} = Y_{it}(e^{u_{it}} - 1) \tag{8}$$

If we plug (8) into (4) in both consecutive periods and use (2), we get:

$$\beta_{it} = \theta_{it} \cdot (e^{u_{it-1}} - 1)/(e^{u_{it}} - 1) \tag{9}$$

This equation states that $\beta_{it} = \theta_{it}$ if, and only if, the log difference between total and reported coronavirus cases (u_{it}) is time invariant, that is when $\Delta u_{it} = u_{it} - u_{it-1} = 0$. Using (2) and (5) and the definition of u_{it} in (7), the previous decomposition in (6) collapses to:

$$\beta_{it} = \beta_{it}^{*} e^{-\Delta u_{it}} \tag{10}$$

This equation shows that the total epidemic curve coincides with the epidemic curve of reported cases when the proportion of undocumented cases does not change over time, i.e., when $\Delta u_{it} = 0$. On the other hand, equation (10) suggests that the epidemic curve of reported cases (i.e. β_{it}) can be estimated using two approaches: i) from an econometric specification of equation (2) that does not provide any information about the relative importance of undocumented cases, as in Orea and Álvarez (2020);[6] or ii) from a stochastic frontier specification of (10) that is able to estimate both the total epidemic curve (β_{it}^{*}) and the temporal changes in the proportion of undocumented cases (Δu_{it}). The latter empirical strategy is developed in detail in the next section. In a nutshell, this strategy implies estimating the epidemic curve of total cases using a stochastic frontier specification of the model where the undocumented cases are proxied using a one-sided random term in the same fashion as firms' inefficiency in production economics. The two partial epidemic curves (i.e., the epidemic curves of reported and undocumented cases) can be obtained once the epidemic curve of total cases has been appropriately adjusted using the estimated proportions of undocumented cases that appear in (9) and (10).

3. Frontier specification of our epidemic curves

3.1. Frontier specification

This section discusses estimation of the epidemic curve of reported case using a stochastic frontier model, an econometric specification widely used in production economics to measure firms' efficiency. The stochastic frontier analysis approach can be used to control for the existence of undocumented coronavirus cases because these cases are not observed by the econometrician and the reported cases are always lower than the total number of COVID-19 infections. This is illustrated in Figure 2, where we have simplified our previous figure by dropping the two partial epidemic curves. This figure shows that the total epidemic curve can be viewed as a function that envelops the observed number of coronavirus cases from above. The gap between Y^{*} and Y is the number of undocumented cases, which never takes negative values. The stochastic frontier analysis approach uses one-sided random terms to control for

[6] This simple empirical strategy might provide biased results as it ignores the potential correlation with the undocumented cases, which constitute an omitted variable in this analysis.

non-negative (or non-positive) unobserved variables, such as firm inefficiency in production economics.

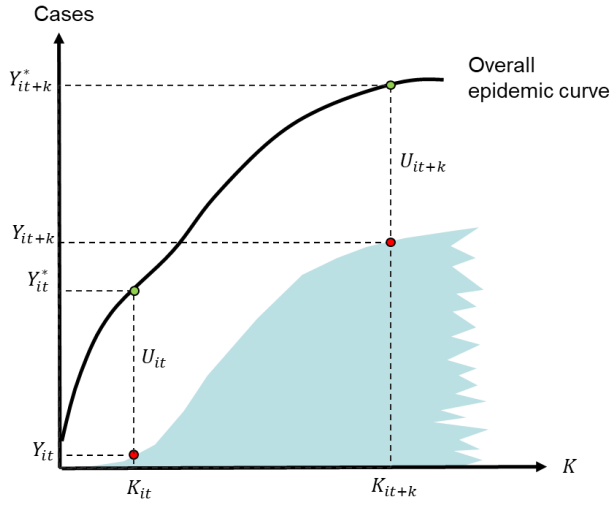

Figure 2: Overall epidemic curve and undocumented cases

As $ln\beta_{it} = \Delta lnY_{it}$ by definition, the stochastic frontier model that is finally estimated can be obtained once we take natural logarithms in (10) and add a traditional noise term:

$$\Delta lnY_{it} = ln\beta^*_{it}(\cdot) + v_{it} - \Delta u_{it} = ln\beta^*_{it}(\cdot) + \varepsilon_{it} \quad (11)$$

where $ln\beta^*_{it}(\cdot)$ is a function of a set of covariates determining the temporal evolution of total coronavirus cases. The idiosyncratic feature of our frontier specification of the model is the existence of two random terms. The first one is the traditional noise term (v_{it}) capturing random shocks, measurement or specification errors and other unobservable variables not correlated with the set of explanatory variables determining the rate of growth of coronavirus cases. The second random term is the difference of two one-sided random terms and captures changes over time in the proportion of undocumented cases (Δu_{it}).

Our empirical strategy thus relies on three assumptions: i) the epidemic nature of this disease can be best represented by a total epidemic curve, regardless of whether researchers observe all COVID-19 cases or not; ii) the unobserved cases can be proxied using a one-sided random term in the same fashion as firm inefficiency in production economics; and iii) the proportion of undocumented cases varies over time during the evolution of the epidemic.

Our *epidemic* frontier model in (11) looks similar to a (panel) stochastic *production* frontier model. It is common in this literature to estimate the following model in levels:

$$lnY_{it} = \alpha_i + f(X_{it}, \beta) + v_{it} - u_{it} \quad (12)$$

where the subscript *i* stands for firm, X_{it} is a vector of exogenous production drivers, β is a vector of technological parameters, v_{it} is a noise term capturing production shocks, and u_{it} is a non-negative random term capturing firm inefficiency. α_i is a firm-specific intercept aiming to capture characteristics that affect firms' production but that are unobserved or omitted variables. Estimation of the model in (12) using the so-called True Fixed Effects (TFE) model

introduced by Greene (2005)[7] is not easy due to the incidental parameter problem.[8] Wang and Ho (2010) solve this problem using temporal transformations of (12). If we take first differences in equation (12) to remove the time-invariant firm-specific effects, we get:

$$\Delta lnY_{it} = \Delta f(X_{it}, \beta) + v_{it}^{*} - \Delta u_{it} \tag{13}$$

where $v_{it}^{*} = \Delta v_{it}$ follows a (multivariate) normal distribution. The production frontier model in (13) is similar to our epidemic frontier model in (11). There are, however, two main differences. First, while $\Delta f(X_{it}, \beta)$ can be negative in a production economics setting, we need to impose the theoretical restriction $ln\beta_{it}^{*} \geq 0$ due to the cumulative nature of Y_{it}. Second, while the production frontier function represents a "technology" (i.e. an unknown combination of production processes), our frontier represents an underlying epidemic process that involves both confirmed and undocumented cases.

In order to estimate the above model using ML, we are forced to choose a distribution for both the noise term (v_{it}) and the one-sided random term capturing the proportion of undocumented cases (u_{it}). In what follows, we discuss the distribution of v_{it}, the distribution of u_{it}, and the likelihood function.

3.2. Distribution of the noise term

We have added a noise term (v_{it}) in equation (11) in order to directly capture measurement errors in the rate of growth of coronavirus cases. As is customary in the stochastic frontier literature, we assume that the v_{it}'s are independent of the u_i's. If we next assume that v_{it} is independently distributed over time and follows a normal distribution with zero mean, the noise vector $v_i = (v_{i1}, \dots, v_{iT})$ will follow a multivariate normal distribution with a diagonal covariance matrix. Using the notation from Wang and Ho (2010), the density function of the vector v_i is:

$$g(v_i) = (2\pi)^{-\frac{T}{2}} |\Pi|^{-1/2} exp\left\{-\frac{1}{2} v_i' \Pi^{-1} v_i\right\} \tag{14}$$

where Π is the variance-covariance matrix of v_i. We then assume that the noise vector $v_i = (v_{i1}, \dots, v_{iT})$ follows a multivariate normal distribution with a diagonal but heteroskedastic variance-covariance matrix because the volatility of the rates of growth of reported cases decreases throughout the epidemic development:

$$\Pi = \begin{pmatrix} \sigma_{v1}^2 & 0 & 0 & 0 \\ 0 & \sigma_{v2}^2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \sigma_{vT}^2 \end{pmatrix} \tag{15}$$

This specification of the variance-covariance matrix of v_i differs from that used in Wang and Ho (2010) in two important aspects. On one hand, our noise term is heteroskedastic, whereas it follows a homoskedastic distribution in Wang and Ho (2010). On the other, while we assumed that v_{it} is not autocorrelated over time, the first-differences and within

[7] This estimator treats α_i as fixed parameters. If they are treated instead as time-invariant random variables, we get the so-called True Random Effects (TRE) panel stochastic frontier model.

[8] This problem appears when the number of parameters to be estimated increases with the number of cross-sectional observations in the data. In this situation, consistency of the parameter estimates is not guaranteed even if $N \to \infty$.

transformations carried out by Wang and Ho (2010) to remove time-invariant firm-specific effects introduce negative correlations between two consecutive (transformed) noise terms.

An autocorrelated specification can be obtained if we introduce the noise terms before computing the rates of growth of coronavirus cases, in the spirit of Chudik et al. (2020) and Millimet and Parmeter (2020). Let us rewrite (7) as follows:

$$lnY_{it}^{*} = lnY_{it} + u_{it} - v_{it} \tag{16}$$

where v_{it} is a two-sided error term that now captures non-systematic variations in total coronavirus cases (Millimet and Parmeter, 2020). If we next take natural logarithms in (5) and replace lnY_{it}^{*} and lnY_{it-1}^{*} with (16) evaluated at t and -1 , we get:

$$\Delta lnY_{it} = ln\beta_{it}^{*} + \tilde{v}_{it} - \Delta u_{it} \tag{17}$$

where $\tilde{v}_{it} = \Delta v_{it}$. The new noise term is no longer independently distributed over time. If we assume that v_{it} follows a heteroskedastic normal distribution, the noise vector $\tilde{v}_i = (\tilde{v}_{i1}, \dots, \tilde{v}_{iT})$ follows a multivariate normal distribution with the following variance-covariance matrix:

$$\Pi = \begin{pmatrix} \sigma_{v1}^2 + \sigma_{v0}^2 & -\sigma_{v1}^2 & 0 & \dots & 0 \\ -\sigma_{v1}^2 & \sigma_{v2}^2 + \sigma_{v1}^2 & -\sigma_{v2}^2 & \dots & 0 \\ 0 & -\sigma_{v2}^2 & \sigma_{v3}^2 + \sigma_{v2}^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & -\sigma_{v(T-1)}^2 & \sigma_{vT}^2 + \sigma_{v(T-1)}^2 \end{pmatrix} \tag{18}$$

If v_{it} is homoskedastic as in Wang and Ho (2010), we get the variance-covariance matrix of their first-differences transformed noise term (see their equation 12). It is an empirical question whether specification (15) or (18) of the noise term is better. However, it should be mentioned that estimation of a frontier epidemic model using (18) is more problematic if the panel dataset is not continuous and there are missing observations between $t = 1$ and $t = T$. This happens, for instance, if we drop the observations with zero rates of growth of coronavirus cases that led to convergence problems[9] when maximizing the likelihood functions in most of our estimated models.[10]

3.3. Distribution of u_{it}.

We now turn to the part of the likelihood function related to the proportion of undocumented cases. Estimating (11) is far from straightforward because the distribution of Δu_{it} is generally not known if we assume that u_{it} is independently distributed across provinces and over time (see, for instance, Wang, 2003, and Orea and Álvarez, 2019). To deal with this issue, we follow Wang and Ho (2010) and assume that u_{it} possesses the so-called scaling property so that it can be multiplicatively decomposed into two components as follows:

$$u_{it} = h(z_{it}, \tau) \cdot u_i \tag{19}$$

[9] Their inclusion makes the rates of growth of coronavirus cases extremely volatile, especially at the beginning of the epidemic outbreaks. This extremely high volatility is difficult to capture using the standard distributions for both the noise term (v_{it}) and the one-sided random term capturing the proportion of undocumented cases (u_{it}).

[10] The number of zero rates of growth decreases notably if we use more recent temporal windows (i.e. not centred around the start of the lockdown) to carry out our empirical analysis. For this reason, we will try to deal with this issue in our empirical application by using a temporal window that begins one week later, at the expense of a fall in the number of pre-lockdown observations.

where $h_{it} = h(z_{it},\ \tau) \geq 0$ is a deterministic (scaling) function, z_{it} is a set of undocumented-cases determinants (often labelled as contextual or z-variables), and u_i is a homoskedastic one-sided random variable. For notational ease, we assume hereafter that the panel dataset is balanced in the sense that we have not dropped observations along the epidemic development. The preceding implies that the first temporal difference of u_{it} in (11) can be rewritten as:

$$\Delta u_{it} = (h_{it} - h_{it-1}) \cdot u_i = \Delta h_{it} u_i \tag{20}$$

Note that the distribution of u_i is not affected by the first-differences transformation. This key aspect of their model enabled Wang and Ho (2010) to get a tractable likelihood function for their transformed model. The same applies to our stochastic frontier epidemic model. Consequently, as the density function of $\varepsilon_{it} = (\varepsilon_{i1}, \dots, \varepsilon_{iT})$ has a closed-form, equation (11) can be estimated by Maximum Likelihood (ML), provided that the scaling function h_{it} is not constant. As Wang and Ho (2010) point out, this condition requires that z_{it} contains at least one variable which changes values over time. Obviously, this happens if we include the epidemic time $K_{it} = t - E_i$ as determinant of the proportion of undocumented cases.

Our frontier model in (20) essentially mimics the one proposed by Wang and Ho (2010) to get a tractable likelihood function for their transformed model. It also looks like the specification introduced by Kumbhakar (1990) and Battese and Coelli (1992) except for the first-differencing transformation of the scaling function. In this sense, we have basically replaced $\eta_{it} = e^{-\eta(t-T)}$ in Battese and Coelli (1992, eq. 2) with $\eta_{it} = \Delta h_{it} = e^{\tau z_{it}} - e^{\tau z_{it-1}}$ where z_{it} is a set of undocumented-cases determinants that include a time-trend variable (e.g. t or K_{it}).

3.4. Likelihood function.

For simplicity, we will assume that $u_i \sim N^+(0, \sigma_u)$. We recall that the half-normal distribution of u_i is not affected by the first-differencing transformation of the idiosyncratic one-sided error term, so that $\Delta u_{it} = \Delta h_{it} u_i$ is distributed as a heteroscedastic half-normal. Wang and Ho (2010) showed that the aforementioned assumptions on v_{it} and u_{it} yield the following log-likelihood function for province i:

$$\begin{aligned} lnL_i = &-\frac{N}{2} ln(2\pi) - \frac{1}{2} ln|\Pi| - \frac{1}{2} \varepsilon_i{}' \Pi^{-1} \varepsilon_i \\ &+ \frac{1}{2}\left(\frac{\mu_*^2}{\sigma_*^2} - \frac{\mu^2}{\sigma_u^2}\right) + ln\left[\sigma_* \Phi\left(\frac{\mu_*}{\sigma_*}\right)\right] - ln\left[\sigma_u \Phi\left(\frac{\mu}{\sigma_u}\right)\right] \end{aligned} \tag{21}$$

where Φ is the standard normal cumulative distribution function, $\varepsilon_i = (\varepsilon_{i1}, \dots, \varepsilon_{iT})$, $\varepsilon_{it} = \Delta lnY_{it} - ln\beta_{it}^*(\cdot)$, and

$$\mu_* = \frac{\mu/\sigma_u^2 - \varepsilon_i{}' \Pi^{-1} \Delta h_i}{\Delta h_i{}' \Pi^{-1} \Delta h_i + 1/\sigma_u^2} \tag{22}$$

$$\sigma_*^2 = \frac{1}{\Delta h_i{}' \Pi^{-1} \Delta h_i + 1/\sigma_u^2} \tag{23}$$

where $\Delta h_i = (\Delta h_{i1}, \dots, \Delta h_{iT})$. Consistent parameters estimates can be obtained by numerically maximizing $lnL = \sum_{i=1}^{N} lnL_t$.

3.5. Frontier specifications of epidemiological models: SIR and epidemic-time frontiers

The frontier model introduced above provides a general specification of the models to be estimated in our empirical section. In order to empirically implement these models, we draw on existing epidemiological models to get a better idea of the variables to be included in the

empirical specification and its possible functional form. In particular, we draw on the Susceptible-Infected-Recovered (SIR) specification of Chudik et al. (2020) and discuss how it can be expressed in a frontier setting.

Chudik et al (2020) derive the following second-order non-linear difference equation specification of the SIR model (see their equation 11):

$$\tilde{Y}_{it} = \tilde{Y}_{it-1}^2/\tilde{Y}_{it-2} + \theta[\tilde{Y}_{it-1}\tilde{Y}_{it-2}(1-\gamma) - \tilde{Y}_{it-1}^2] \tag{24}$$

where $\tilde{Y}_{it}$ denotes the true number of infected in province i at time t, θ is the effective transmission rate, and γ is the rate of recovery.[11] If we divide both sides of (24) by $\tilde{Y}_{it-1}$ and take logs, we get:

$$\Delta ln\tilde{Y}_{it} = ln\left[\tilde{Y}_{it-1}/\tilde{Y}_{it-2} + \theta[\tilde{Y}_{it-2}(1-\gamma) - \tilde{Y}_{it-1}]\right] \tag{25}$$

As can be seen, the true rate of growth of coronavirus cases depends on first- and second-order lagged values and their interaction. Ignoring other random errors, Chudik et al. (2020) assume that the ratio of confirmed to true cases at time t can be written as:

$$\frac{Y_{it}}{\tilde{Y}_{it}} = \pi_{it} = e^{-u_{it}},\ u_{it} \geq 0 \tag{26}$$

so that

$$Y_{it}e^{u_{it}} = \tilde{Y}_{it},\ u_{it} \geq 0 \tag{27}$$

where the one-sided term u_{it} in (27) simply measures the gap between the true and confirmed number of cases, such that:

$$ln\tilde{Y}_{it} = lnY_{it} + u_{it} \tag{28}$$

If we use (28) to replace the true number of cases on the left-hand side of (25) with their "observed" counterparts, we get:

$$\Delta lnY_{it} = lnf(q_{it}, \beta) - \Delta u_{it} \tag{29}$$

where

$$f(q_{it}, \beta) = \left[\tilde{Y}_{it-1}/\tilde{Y}_{it-2} + \theta[\tilde{Y}_{it-2}(1-\gamma) - \tilde{Y}_{it-1}]\right]$$

Note that $f(q_{it}, \beta)$ is a function of a set of covariates determining the temporal evolution of *total* coronavirus cases. It depends on the true, but unobserved, number of cases in periods $t-1$ and $t-2$. In order to estimate (29), we can follow Chudik et al (2020) and replace them with their "observed" counterparts. In this case, equation (27) becomes:

$$\Delta lnY_{it} = ln\left[\frac{Y_{it-1}}{Y_{it-2}} \cdot e^{\Delta u_{it-1}} + \theta[Y_{it-2}e^{u_{it-2}}(1-\gamma) - Y_{it-1}e^{u_{it-1}}]\right] - \Delta u_{it} \tag{30}$$

If we assume that the one-sided random term u_t is i.i.d. and follows, say, a half-normal distribution, the distribution of (30) in not known and cannot be estimated using the standard stochastic frontier (SF) estimators.

In order to estimate the SIR-based frontier (30) using standard SF techniques, we need to make some simplifying assumptions. Concretely, we assume that the u-terms inside the brackets

[11] It should be pointed out that we use a wider definition of coronavirus cases than Chudik et al (2020) in equation (24). Our coronavirus cases in time t not only include current infected individuals in time t but also previously infected people that had already recovered or deceased at that time. Although a similar expression can be obtained using our definition of coronavirus cases, we use (24) to facilitate the points set out below.

balance each other out and, as is customary in the production economics literature, we can assume that $lnf(q_{it},\beta)$ is a linear function of lnY_{it-1}, lnY_{it-2} and $lnY_{it-1}lnY_{it-2}$.[12] Whether such an approximation is valid is an empirical question.[13]

The theoretical SIR model suggests that $\tilde{Y}_{it-1}$, $\tilde{Y}_{it-2}$ and their natural logarithm can be accurately approximated using a third-order function of lnK_{it}. Therefore, another empirical strategy that does not require using sophisticated econometric techniques is to estimate (29) under the assumption that $f(q_{it},\beta)$ is a function of lnK_{it}, lnK_{it}^2 and lnK_{it}^3. If the measurement errors associated to replacing $ln\tilde{Y}_{it-1}$ and $ln\tilde{Y}_{it-2}$ with the abovementioned three variables do not disturb the distribution of the (composed) error term, this version of the model can also be estimated using simple SF techniques.

3.6. Simulation exercises

To check the performance of the frontier logged specification of the SIR model (using temporal lags of lnY_{it}) and the approximation of the SIR model (using functions of lnK_{it}), which we label the *epidemic-time* specification, we carry out a simulation exercise where we vary both the panel structure of the data (modifying the cross-sectional and time dimensions) and the parameters of the error terms.

We first generate the true evolution of total coronavirus cases in a representative cross-section unit (province) using the discrete-time SIR model developed by Chudik et al. (2020). Following these authors, we assume in our simulations that the number of days to recovery or death (d) is equal to 14, and that the basic reproduction number (R_0) is equal to 3.[14] The number of individuals (as a fraction of population) that have not yet contracted the disease, have recovered or died, or are still infected in day t are simulated respectively using their equations 5, 7, and 11, assuming no social distancing interventions.[15] Unit-specific values are then obtained by adjusting the theoretical values with simulated values for both the noise term (v_{it}) and the one-sided random term (u_i) capturing the proportion of undocumented cases. While the values of the noise term are simulated using a normal distribution, the simulated values for the one-sided random term are obtained using a half-normal distribution.

In the most general form of the model, the random error terms have a heteroskedastic specification where they depend on the epidemic time of each cross-sectional unit (province). This general model can be expressed as follows:

$$\Delta lnY_{it} = f(q_{it},\beta) + v_{it} - (e^{\eta_u K_{it}} - e^{\eta_u K_{it-1}})u_i \tag{31}$$

$$v_{it} \sim N(0, e^{ln\sigma_v + \eta_v K_{it}}) \tag{32}$$

$$u_i \sim N^+(0, e^{ln\sigma_u}) \tag{33}$$

where the $f(q_{it},\beta)$ in (31) is either the SIR specification or its epidemic-time approximation using first and higher-order terms of lnK_{it}. The noise term is specified as a function of the

[12] Squares of both lnY_{t-1} and lnY_{t-2} can also be included if a Translog specification is preferred.

[13] The model may be biased not only because we are ignoring u-terms but also because the lagged values of reported cases might be correlated with the time-invariant part of the error term capturing the proportion of undocumented cases (u_i). This could occur if undocumented (asymptomatic) cases facilitate the dissemination of COVID-19 and thereby increase the reporting rates. This is a hotly-debated issue in the epidemiological literature and is yet to be resolved.

[14] That is, we assume that the recovery rate ($\gamma = 1/d$) is equal to 1/14 and that the rate of transmission ($\beta = \gamma R_0$) is equal to 3/14.

[15] Total number of individuals are obtained assuming a population size equal to 1,000 inhabitants.

epidemic time where the parameter η_v can be assigned negative values to reflect decreasing volatility of rates of growth of reported cases over time. The undocumented cases, captured by the one-sided error time, u_{it}, are modeled as a function of the epidemic time through the scaling function, while the parameter η_u can be assigned negative values to capture reductions over time in the proportion of undocumented cases.

The simulations are carried in all cases with a total of 1,000 observations but with two different panel data structures where N (number of cross-section units) and T (number of time periods) are varied. The panel structures are (i) $(N = 20; t = 50)$ and (ii) $(N = 40; T = 25)$ and the simulation results are presented in separate tables for each structure (Tables 1 and 2). In each table, Model 1 refers to the epidemic-time model where $q_t = (lnK_{it}, lnK_{it}^2$ and $lnK_{it}^3)$ and Model 2 corresponds to the SIR specification where $q_t = (lnY_{it-1}, lnY_{it-2}, lnY_{it-1}lnY_{it-2})$.

[Insert Table 1 here]

[Insert Table 2 here]

The first block of columns in each table shows the parameter settings of the random error terms used to generate the data and the corresponding true average multiplication factors (MF). Changes are introduced to the parameter settings used to generate the data by modifying the settings of $ln\sigma_v$, $ln\sigma_u$, η_u and η_v. The remaining columns contain the results from the epidemic-time and SIR specifications where the models are estimated in every case with the parameter specifications $ln\sigma_v = -3$ and $\eta_v = 0$. We compare the parameter estimates of these models with the true parameters used to generate the data, paying special attention to the accuracy of the estimate of the distribution of the parameter underlying the multiplication factor, $ln\sigma_u$.

We begin by noting that the first two columns in the blocks of results compare the performances of the non-frontier $(u = 0)$ and frontier $(u \geq 0)$ specifications of each model based on their R^2 statistics, and it can be seen that in all cases the frontier specification performs better. Consequently, in the discussion of results that follows, the frontier specification is used. It should also be noted that the R^2 of the SIR models are higher in all cases, both for non-frontier and non-frontier specifications. This is an expected results because the cross-sectional information contained in lnY_{it} is much richer than contained in lnK_{it}.

For each panel data structure and for each estimated model, four sets of simulations are presented, each set containing two different specifications of $ln\sigma_u$ in the data generating process: simulations 1a, 2a 3a and 4a set $ln\sigma_u = 1$ whereas simulations 1b, 2b, 3b and 4b set $ln\sigma_u = 0.5$.

We start with Table 1, where the panel structure is $N = 20$ and $T = 50$ and focus first on the results from Model 1. The first set of simulations (1a and 1b) show that the model performs well, with the estimated values of $ln\sigma_u$ close to their true values and very high correlations (> 0.99) between the estimated and true u. Similarly, the estimated mean squared error (MSE) between the estimated and true u is small (0.001 and 0.002).

The second set of simulations (2a and 2b) illustrate what happens when heteroskedasticity is introduced into the symmetric error term ($\eta_v = -0.1$) in the data generating process and this is not accounted for by the estimated model which, we recall, is estimated assuming that the symmetric error term is homoskedastic ($\eta_v = 0$). In this case it is clear that the model performs very poorly, with estimated values of $ln\sigma_u$ substantially higher than the true values. This is reflected in much larger MSEs and smaller correlations between the true and estimated u than the previous set of simulations. Thus, if we suspect that the symmetric error term is heteroskedastic then we should make sure to model this.

In simulations 3a and 3b, the temporal decay of the one-sided error term, u, is reduced compared to simulations 1a and 1b as reflected by the change in η_u to -0.1. In the final set of simulations we check what happens when the true variance of the symmetric error term is increased, this time by increasing the value of $ln\sigma_v$. Note that in these final two sets of simulations the true symmetric error term is homoskedastic ($\eta_v = 0$). As can be seen, the estimated models perform relatively poorly, substantially underestimating the true $ln\sigma_u$ in all cases and with large MSEs. The model performs particularly poorly when the symmetric error variance is large.

Regarding the performance of Model 2 (SIR specification) for this panel data structure, a comparison with Model 1 shows that Model 2 generally provides a better goodness-of-fit, with higher values of R^2 in every case. It can be seen that whereas Model 1 tended to underestimate $ln\sigma_u$ (the exceptions being simulations 2a and 2b), Model 2 tends to overestimate it (with the exception of simulations 4a and 4b), thereby overestimating the unconfirmed cases. As with Model 1, this overestimation is particularly pronounced when the data are generated with heteroskedasticity in the symmetric error tem (simulations 2a and 2b). It can also be noted that Model 2 also performs much worse at predicting the parameter $\hat{\eta}_u$.

Turning to Table 2, where the panel structure is $N = 40$ and $T = 25$, we see that Model 1 again underestimates $ln\sigma_u$, this time in all cases, including that when the data are generated with heteroskedasticity in the symmetric error term. In this particular case, the model performs much better than under the previous panel structure, implying that the greater heterogeneity in the cross section counteracts to a certain extent the effect of the unmodelled heteroskedasticity in the symmetric error term. As for the SIR specification, $ln\sigma_u$ is underestimated in two sets of simulations (1a and 1b, and 4a and 4b) and overestimated in the other two. With heteroskedasticity in the symmetric error term in the data generating process, the model performs much better than in the previous panel structure, as happens with Model 1. Note that the MSEs and correlations between the true and estimated u values are relatively less reliable in Model 2 than Model 1.

Overall, the simulation exercise throws up some interesting results which serve as a guide towards designing an appropriate empirical specification of our frontier model. Generally, the SIR specification provides a better goodness of fit than the epidemic-time specification because lnY_{it} exhibits greater cross-sectional heterogeneity than lnK_{it}. However, the SIR specifications tend to overestimate the $ln\sigma_u$ parameter and also tends to provide poorer estimates of the parameter η_u, leading to it overestimating the (proportion of) unreported cases. The epidemic-time specification, on the other hand, tends to slightly underestimate the proportion of unreported cases but provides relatively more accurate estimates of $ln\sigma_u$ and η_u. Finally, in the panel structure with the smaller cross-section, both models perform particularly poorly when they do not take into account heteroskedasticity in the symmetric error term. When the cross-section dimension is increased the estimates are much more accurate. In any case, these point to the appropriateness of a modelling the symmetric error term as heteroskedastic.

4. Empirical illustration

4.1. Sample and data

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.

The daily evolution of laboratory-confirmed COVID-19 cases in the Spanish mainland provinces was collected manually by the authors from the official press releases of the Spanish regional governments, the Ministry of Health and Wikipedia. These information sources had to be consulted to extend backwards the provincial data published by *Datadista* in GitHub, under a free license. GitHub extracts their data from a variety of documents published by the Ministry of Health but only published data from March 13th on.[16] For the 28th of March onwards we collected the data directly using *RTVE Flourish*.[17] We used the regional online data released by the Ministry of Health[18] and the province-level dada released by the Spanish regional governments to correct typos and lack of information on coronavirus cases in some provinces.

We do not show the temporal evolution of reported coronavirus cases in each province for reasons of space but they can be found in Orea and Álvarez (2020). Instead, in Figure 3 we show the onset epidemic dates of each province, which determines the values of the epidemic times. A feature worth highlighting is the relatively large dispersion of onset dates across provinces. This feature is crucial for the estimation of the epidemic-time version of (26) because we need observations with both small and large epidemic times to appropriately estimate the parametric function of lnK_{it}, especially before the lockdown implementation date.

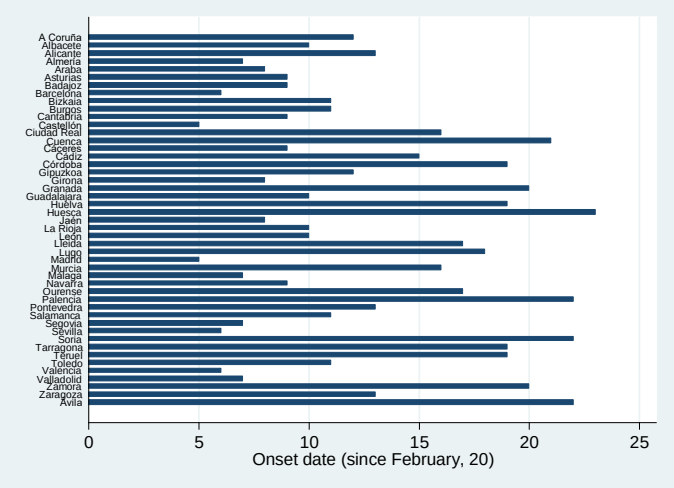

Figure 3: Epidemic onset dates

Rather than trying to directly explain (predict) the number of reported cases, we use the rates of growth of reported coronavirus cases to estimate (11) as we have found that this variable is

[16] See https://github.com/datadista/datasets/tree/master/COVID%2019.

[17] See https://app.flourish.studio/visualisation/1451263/.

[18] See https://covid19.isciii.es/.

stationary.[19] Figure 4 shows the boxplots of the rates of growth of reported cases by epidemic time. Two features are evident from this figure. First, the rates of growth of reported cases are much larger at the beginning of the epidemic than when the epidemic has advanced. That is, our dependent variable tends to decrease over the epidemic time. Second, the volatility is much larger when K_{it} is small, and declines as K_{it} increases. This calls for a time-varying heteroskedastic specification of our symmetric error term.

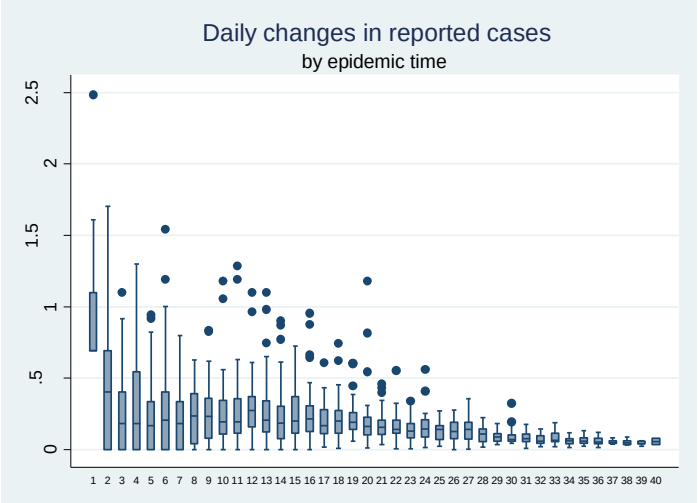

Figure 4: Rates of growth of reported cases

It should be mentioned here that zero rates of growth often appear at the beginning of outbreaks, where our dependent variable looks like a count variable with the observations taking a small range of non-negative integer values. Once the slope of the epidemic curve increases, our dependent variable no longer has this feature. Allowing for zero rates of growth tended to produce convergence problems when maximizing the likelihood functions. For this reason, we estimate our epidemic models dropping the observations with zero rates of growth.[20]

4.2. Parameter estimates

Table 3 shows the parameter estimates of several epidemic-time specifications of equation (11). That is, the four specifications in this table use a third-order function of lnK_{it} to capture the true temporal pattern of the virus epidemic. As the likelihood function of these models has a closed form, they have all been estimated by ML. The first two specifications assume that the epidemic curve of total coronavirus cases (i.e., $ln\beta_{it}^{*}$) is a linear function of a set of covariates, whereas the last two specifications assume that $ln\beta_{it}^{*}$ is an exponential function in order to impose the theoretical restriction $\beta_{it}^{*} \geq 1$. The non-frontier models assume that $\Delta u_{it} = 0$, thereby ignoring the one-sided random term that appears in equation (11), which is equivalent to assuming that the proportion of undocumented cases does not change over time. These non-frontier models therefore impose the strong assumption that the epidemic curves of both total

[19] A Harris-Tzavalis (1999) unit-root test allows us to reject that ΔlnY_{it} contains unit roots. The value of the statistic is 0.0338 with a *p*-value equal to zero.

[20] We often achieved convergence when estimating non-frontier specifications of the model. We found in these cases that only the initial temporal patterns tended to be biased upwards. We will return to the issue of dropping zero-growth in our frontier specifications of the model in the robustness analyses section.

and reported coronavirus cases coincide (see equation 10). The frontier models relax this assumption by adding the first difference of a one-sided error term that can be multiplicatively decomposed into an exponential scaling function (that is, we assume that the scaling function that appears in equation (12) is $h_{it} = e^{z_{it}'\tau}$) and a homoskedastic half-normal random variable.

[Insert Table 3 here]

All models include a day-of-the-week effect (not reported) that aims to capture reporting lags by regional and national governments. They also all include a dummy variable D_t that takes the value 1 from the 14th of March, 2020, the day marking the imposition of most of the coronavirus control measures by the Spanish government. The coefficient of this dummy variable allows us to test whether the Spanish lockdown and other public control measures implemented around the 14th of March were able to attenuate the spread of the virus within each province.[21] Notice that our model specification looks like a Difference-in-Difference model where we compare an outcome variable before and after treatment (a policy measure). Although the lockdown of the population in Spain was implemented on March 14th in all provinces, there were substantial differences in the evolution of the epidemic in each province on that date. Therefore, our identification strategy is based on the relatively large dispersion of epidemic onset dates across provinces and the onset dates being orthogonal to the lockdown implementation date.

Following the scant epidemiology literature that controls for spatial spillover effects, we use a spatial lag of X specification (SLX) to measure the propagation effect of mobility of people across provinces. In particular, we include $W_i lnK_t$ as an epidemic frontier driver, where lnK_t is a $Nx1$ vector of epidemic times of the Spanish provinces, and W_i is a $1xN$ spatial weight vector where the weights measure the degree of mobility (connectivity) between provinces. We follow Giuliani et al. (2020) and Gross et al. (2020) and use a contiguity or binary W_i vector, where the weights equal one for adjacent units and zero for non-bordering units.[22] Therefore, we assume that $ln\beta_{it}^*$ depends on the epidemic time of neighbouring provinces. We have selected the epidemic time to capture the potential propagation effects between provinces for two reasons. First, this variable is exogenous by construction. Second, Vega and Elhorst (2015, p. 342) suggest taking the SLX model as the point of departure because this specification is not only the simplest specifications but also more flexible in modelling spatial spillover effects than other specifications.[23]

The specification of both random terms is also common to all models. On the one hand, all models have been estimated using a heteroskedastic specification of the noise term because the volatility of rates of growth of reported cases decreases throughout the evolution of the epidemic. In particular, we assume hereafter that the logarithm of the standard deviation of v_{it} depends on the logarithm of K_{it}. On the other hand, we recall that our empirical strategy relies

[21] It is worth mentioning that the third-order function of lnK_{it} captures the temporal pattern of the virus epidemic, conditional on D_t. In other words, the epidemic curve associated to this function can be interpreted as our *as if* scenario with no control measures.

[22] Other spatial specifications based on students' regions of origin, high-speed railway connectivity, and the tourist habits of city-residents and their regions of origin were used in Orea and Álvarez (2020). We find very similar results using alternative spatial specifications.

[23] Our spatial SLX specification does not distinguish between reported and undocumented propagation across provinces. A SAR specification with $W_i lnY_t$ and $W_i u_t$ allows us to deal with this issue. However, estimating this model is far from simple because the distribution of $W_i u_t$ is generally not known if u_{it} is independently distributed across provinces, as assumed above. As estimating this model presents important methodological challenges, we leave an examination of this issue for future research.

on a time-varying proportion of undocumented cases during th evolution of the epidemic (see Li et al., 2020). In order to capture temporal changes in u_{it}, we assume that the scaling function depends on two time-varying contextual variables: i) the epidemic time of each province (K_{it}), in the same fashion as Battese and Coelli (1992); and ii) the logged epidemic time of neighbouring provinces ($W_i lnK_t$), because we believe that the mobility of people across provinces might also have a significant effect on the proportion of undocumented cases. Finally, we also interact our lockdown dummy variable D_t with both K_{it} and $W_i lnK_t$ in order to examine whether the Spanish lockdown and other control measures (such as an increase in testing) have also reduced the proportion of undocumented cases.

Table 3 presents the results of the epidemic-time version of the models that use a third-order function of lnK_{it} to capture the temporal evolution of coronavirus cases. The intercepts estimated in the *linear* models are close to unity, indicating that the initial rates of growth of coronavirus cases are relatively large. The *exponential* models yield much larger initial growth rates, a result that might explain why all the coefficients of the third-order function of lnK_{it} are statistically significant using this specification. In contrast, we do not find significant lnK_{it} coefficients using the linear specification, a result that seems to (incorrectly) suggest that the rates of growth of coronavirus cases do not change during the epidemic. Figure 4 suggests, however, that these rates of growth decrease rapidly in the early stages of the epidemic.[24] This feature is better captured by the exponential model as the negative large coefficient of lnK_{it} found using this specification indicates that these growth rates rapidly decreased a short time after the beginning of the epidemic. Moreover, the previous result, together with the positive and negative coefficients found respectively for lnK_{it}^2 and lnK_{it}^3, is consistent with the traditional S-shaped epidemic curves. For all these reasons, the exponential specifications of our epidemic curve are the preferred ones.

Another key result of our empirical exercise is the positive and statistically coefficient found for the spatially-lagged variable, $W_i lnK_t$. This indicates that the rate of growth of COVID-19 cases in one province depends on the development of the epidemic in other provinces. In other words, two provinces with similar epidemic histories would evolve differently if one is close to one of the epicentres of the coronavirus in Spain and the other is far from these epicentres.[25] Notice that we have interacted D_t with $W_i lnK_t$. This implies that the coefficient of $W_i lnK_t$ actually measures propagation effects *before* the implementation of the Spanish lockdown. The positive value found for this coefficient therefore provides evidence supporting the belief that the exodus of city residents and students living in the epicentres of the Spanish coronavirus crisis that left cities to spend their confinement in their family and vacation homes - located in provinces that still did not have coronavirus cases or that were in the early stages of development of their coronavirus epidemics - encouraged spread the virus across the country.

On the other hand, the coefficient of $W_i lnK_t \cdot D_t$ is negative and statistically significant, indicating that the lockdown has been quite effective in preventing the propagation of the coronavirus *between* provinces. This result seems to confirm the huge reduction in people's mobility found by Google (2020) in its report for Spain. Mobility trends for workplaces and public transport hubs (such as subway, bus and train stations) decreased by 63% and 84% respectively from the 29th of February to the 11th of April.[26]

[24] This figure also shows a flattening in mid-stages of the epidemic, and very small rates in later stages of the epidemic. In other words, Figure 2 indicates that the epidemic curve has the traditional S-shaped form.

[25] Gross et al. (2020) find a strong correlation between the number of infected individuals in each province and the population migration from Hubei, the main epicentre of the Chinese epidemic, to each province.

[26] Interestingly, mobility trends for places of residence increased by 26% in the same period.

Another issue is whether the lockdown has been effective in reducing the propagation of the virus *within* each province. This within-province impact of the Spanish lockdown can be examined using the estimated coefficient of D_t. As $W_i lnK_t$ is measured in deviations with respect to the post-lockdown sample mean, the coefficient of D_t can be interpreted as an average effect. We find a negative and statistically significant effect of the Spanish lockdown on the rates of growth of coronavirus cases, regardless of whether we use linear or exponential specifications for the epidemic curve.[27] The estimated effect of the Spanish lockdown on the rates of growth of coronavirus cases is likely to be biased upwards if we use a linear specification because, using this specification, the set of epidemic time variables is not able to capture the observed decline in the rates of growth of coronavirus cases. This does not occur if we use an exponential specification, which is able to produce the traditional S-shaped epidemic curve. Indeed, while the average effect that is estimated using the linear frontier model is quite large (13.7%), the exponential frontier model produces lower effects of the lockdown on the rates of growth of coronavirus cases (6.8% on average). In summary, these results allow us to conclude that the lockdown has been effective in both preventing the propagation of the coronavirus between provinces and in attenuating the propagation of the virus within each province. In other words, we find that the Spanish lockdown, together with other control measures, has flattened the epidemic curves of all provinces.

We now focus our discussion on the distribution of both random terms. As expected, we find that the standard deviation of the noise term decreases with the logarithm of K_{it}. Regarding the one-sided random term, we find that the coefficients of the scaling function are negative and most of them are statistically significant. This suggests, again as expected, that the proportion of undocumented cases decreases over time. Moreover, we find that the decline in u_{it} is even larger after the lockdown, a result that is in line with Li et al. (2020). Also worth noting are the estimates of the standard deviation of u_i in both the linear and exponential specifications. This is a critical parameter because it conditions the estimated proportion of coronavirus cases that have not been detected by the regional health systems in all Spanish provinces. Despite the relatively modest (logged) standard deviation of u_i that appears in Table 1, we find very different under-reporting rates across the Spanish provinces. The large temporal and cross-sectional heterogeneity in (under-)reporting rates found in our empirical application is one of the contributions of the paper as the previous epidemiological literature often relies on common rates.

The aforementioned heterogeneity will be examined in more detail later, after we discuss the average rates of both reported and undocumented cases and multiplication factors found using our preferred model. We find that the average under-reporting rates ($UR = U/Y^*$) using the exponential frontier model is 57.8%, implying that the average reporting rate ($RR = Y/Y^*$) is 42.2%[28]. The latter percentage lies between the two reporting rates provided by Li et al. (2020) in their study of the Chinese coronavirus epidemic. If we use our estimated average to compute multiplication factors ($MF = Y^*/Y$), we find that the total number of cases that our frontier exponential model predicts is on average more than twice the observed number of cases (i.e. $MF = 2.38$). As such, our average multiplication factor is slightly larger than that obtained by Millimet and Parmeter (2020) for a sample of 63 countries using a non-spatial and non-autoregressive frontier approach. Our findings are close the assumption made by Chudik et al.

[27] One might argue that this effect can only be identified one or two weeks after the implementation of the national lockdown. This is likely true, but we should keep in mind that the social distancing measures, local lockdowns, and closures of schools and universities were implemented before the national lockdown.

[28] Much larger under-reporting rates are obtained if we estimate the model partially using NLLS (discussed in the next section). For this reason, the above average values can be viewed as a lower bound of the true values.

(2020) who specifically account for underreporting using a multiplication factor of two, which they derive from data on the number of asymptomatic individuals aboard the Diamond Princess cruise ship.

Regarding the temporal path of reporting rates, Figure 5 shows the province-specific reporting rates by epidemic time computed using our preferred exponential frontier model. Several comments are in order regarding this figure. First, we observe that all rates tend to increase throughout the evolution of the epidemic because we have found before that u_{it} tends to decline over time. Second, the sample mean varies from 25.3% to 52.5%. These averages reveal that the multiplication factor is on average close to 4 at the beginning of the epidemic and close to 2 at later stages.[29] Third, the minimum RR values suggest that there are (many) provinces with very low reporting rates, and hence extremely large multiplication factors, especially at the very beginning of their epidemic episodes. In this sense, our estimated reporting rates are in line with Li et al. (2020), who also find very low reporting rates (14%) before the implementation of the Chinese travel restrictions.[30]

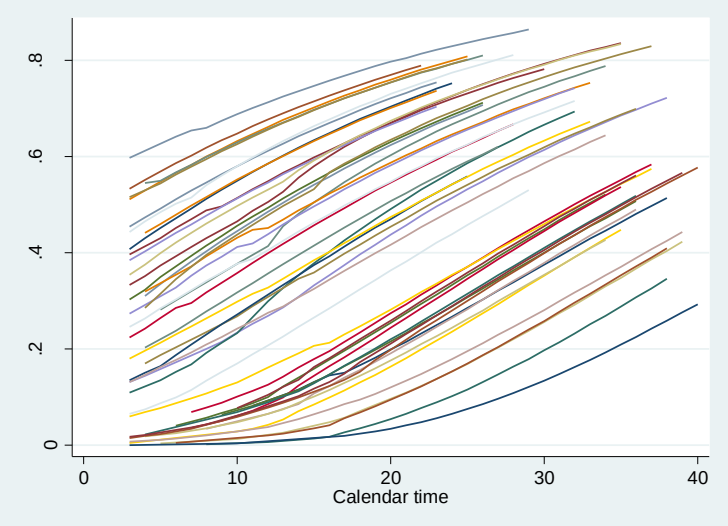

Figure 5: Temporal evolution of reporting rates

Figure 5 shows the large variety of reporting rates in Spain but not the geographical distribution of reporting rates across the Spanish provinces. This is shown in Figure 6. As the reporting rates vary over time, we have depicted this map using the provincial reporting rates evaluated at the epidemic time 20. Figure 6 seems to suggest the existence of two groups of provinces, one with relatively large reporting rates and the other with relatively low reporting rates. Figure 6 shows that most, but not all, of the provinces with small reporting rates are located in the regions of Castilla-León, Extremadura, and Valencia, and the two main epicentres in Spain (Madrid and Barcelona). The multiplication factors in these provinces (not shown) are on average close to 8. The largest reporting rates are found in coastal Andalucía and several provinces located in the Iberian and Pyrenees mountain ranges. Consequently, their

[29] If we use individual reporting rates to compute individual multiplication factors, we get values on the order of two or three digits, with a mean value of 8, which are consistent with the large attack rates (i.e. proportions of infected people) found for Spain by Flaxman et al. (2020) in their study using 11 European countries.

[30] The fraction of all infections that were documented after the travel restrictions was estimated to be 65%, a slightly larger reporting rate than that found in our paper after the implementation of the Spanish lockdown.

multiplication factors are much smaller than those computed for the previously-mentioned provinces (close to 1.7 on average).

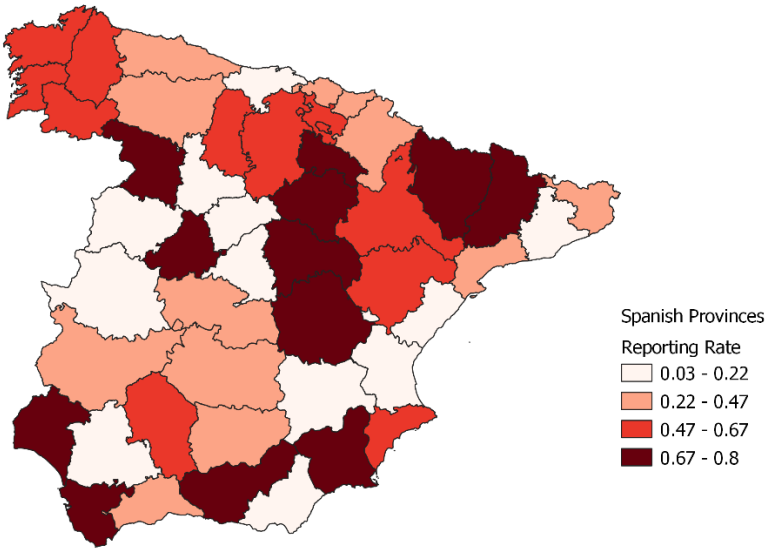

Figure 6: Geographical distribution of reporting rates

4.3. Robustness analyses

In this section we provide some extensions to the base model and robustness checks. First, we compare our previous results with those obtained using a specification inspired in the SIR theoretical epidemic model, presented in Section 3.3, where we replace the third-order function of lnK_{it} with the first and second-order lagged values of lnY_{it} and their interaction. Second, we extend our model by introducing additional variables. An appealing feature of both epidemic-time (lnK_{it}) and SIR-based specifications is that they can be estimated using epidemic-type data only, i.e., the rates of growth of coronavirus cases depend in our models on own and neighbours' epidemic times, lagged cases of COVID-19, date of implementation of control measures, etc. However, this does not preclude adding other covariates. We take advantage of this to explore the influence of a series of socio-economic and environmental variables on both the epidemic frontier and (through the scaling function) the proportion of under-reported cases. Our third analysis of robustness has to do with the distributional assumptions made to obtain ML estimates. A common criticism levelled against the use of stochastic frontier models is that they rely heavily on distributional assumptions. As an alternative, we use a non-linear least squares (NLLS) estimator which does not rely so heavily on distributional assumptions and compare it to our models estimated by ML. A fourth analysis aims to examine the effect of different temporal "windows" when estimating our epidemiological frontier model. A final robustness analysis has to do with the specification of the variance-covariance matrix of our noise term. We assume here that the noise term capturing measurement errors in the rate of growth of coronavirus cases is autocorrelated over time, in the same fashion as Wang and Ho (2010).

4.3.1. Alternative specifications: SIR-based models

We compare the results obtained using the SIR-based specification presented in Section 3.3 with those of our epidemic-time (lnK_{it}) model. The results from the different specifications of the SIR-based model are presented in Table 4.

[Insert Table 4 here]

As in the simulation exercise presented in Section 3.3., we find that the SIR-based specifications provide a better goodness-of-fit in all cases than the epidemic-time models based on lnK_{it}. There are two plausible explanations for this. The first is that the number of reported cases is a cumulative and non-stationary variable, so that the lagged values of lnY_{it} are highly correlated with the rate of growth of reported cases. The other, and more compelling reason, however, has to do with the large cross-sectional heterogeneity of lnY_{it} compared to lnK_{it}, which implies that the cross-sectional differences in lnY_{it} are more informative than the cross-sectional differences in lnK_{it}.

As with the epidemic-time models, we find a positive and statistically coefficient for $W_i lnK_t$ in all the SIR-specification models, indicating that the mobility of the people across provinces did clearly spread the virus across the country. We find a negative and statistically significant coefficient for the interaction of $W_i lnK_t$ with the lockdown dummy variable, D_t, in line again with the epidemic-time specification. Interestingly, the estimated values are larger in absolute terms than those obtained in the corresponding epidemic-time models in Table 3. This seems to suggest that the lockdown has been even more effective in preventing the propagation of the coronavirus *between* provinces using the SIR specification. In contrast, the *within*-province impact of the Spanish lockdown captured by the estimated coefficient of D_t is smaller than in the epidemic-time models. Overall, both the epidemic-time and SIR-based specifications suggest the existence of significant spatial spillovers and provide evidence that the Spanish lockdown and control measures implemented around the 14th of March were effective in reducing the propagation of COVID-19 within and between provinces.

Regarding the two random terms, in the SIR-based models we find a decreasing standard deviation for the noise term, as occurred with the epidemic-time models. The parameter estimates of the scaling function, on the other hand, differ notably from those obtained in the epidemic-time models. For the linear SIR-based model we find no statistically significant coefficients for the time-varying variables of the scaling function, implying that this specification poorly estimates the parameter measuring the standard deviation of u_i. A comparison with the corresponding linear model using the epidemic-time specification is instructive: whereas 65.1% of total cases were found to have been under-reported on average using the linear epidemic-time specification, this rises to 96.6% using the linear SIR-based specification. The flip-side of this that the linear SIR-based model provides extremely low reporting rates (close to 3%). This is not plausible as it is equivalent to an average multiplication factor close to 30. The linear epidemic-time model, in contrast, provides more reasonable rates for reported coronavirus cases (close to 35%). This result seems to confirm the findings from our simulation exercise, where we found that the SIR-based models tend to overestimate the standard deviation of the one-sided random term, and thereby the proportion of undocumented cases.

Whereas the linear SIR-based and epidemic-time models provide very different average reporting rates, the exponential specification of the SIR-based model provides quite similar average reporting rates to its epidemic-time equivalent (40.6% and 42.2%, respectively). The exponential form of the SIR-based frontier epidemic curve therefore tends to attenuate the bias in the estimation of the one-sided error term. This suggests that exponential rather than linear

epidemic specification should be used when estimating a frontier SIR-based model, despite our finding that they provide slightly worse goodness-of-fit than the linear specifications.

4.3.2. Additional variables: socio-economic determinants

Stojkoski et al. (2020) point out that a multitude of social, demographic and economic criteria, aside from the biological and epidemiological factors, influence the extent of the spread of the coronavirus disease through the population, as evidenced during this first wave of the pandemic.[31] To examine this issue, we have estimated our preferred model, namely the exponential epidemic-time specification, by adding, one at a time, a series of socio-economic variables. The introduction of crucial socio-economic determinants not only provides an estimate of their potential impact but may also offer guidance for future policies aimed at preventing the emergence of epidemics.[32] These are introduced firstly as determinants of the one-sided error term (u) measuring the proportion of under-reporting cases, and secondly as drivers of the frontier epidemic curve that captures the overall evolution of the pandemic. Table 5 provides the parameter estimates of the new variables once they have been introduced into the model one at a time.[33]

[Insert Table 5 here]

The socio-economic environment is measured through three variables: the provincial GDP per capita and the shares of the services and agricultural sectors in total provincial employment. The demographic structure is measured using the following variables: population size (in logs), population density (people per squared km, in logs), and three population age variables (percentage of population aged 15-24, percentage of population aged 25-65, and percentage of population aged over 65). As there is an active debate regarding the influence of the natural environment, we have also included two weather variables (temperature and rainfall) that were available at provincial level.

Looking first at the results for the scaling function, which show the influence of these variables on the proportion of under-reporting cases, it can be seen that none of them has a significant effect. While the lack of significance suggests that under-reporting rates are unaffected by these variables, an alternative explanation may lie in the fact that the random u-term we are modelling here is time-invariant. Although their coefficients are not statistically significant, some of the estimated signs are worth mentioning, especially the positive coefficients of youth and middle-age population, and the negative sign of elderly people.

Turning to their effect on the overall epidemic curve, most of the demographic and weather variables do not also have a significant frontier effect. We do find, however, that the most-populated provinces have had more intensive coronavirus epidemics, most likely due to agglomeration of individuals and the fact that the use of public transport is more prevalent in these provinces. We also find that the COVID-19 epidemic is more intense in provinces with

[31] These authors identify a total of 30 potential socio-economic factors, including healthcare infrastructure, societal characteristics, economic performance, demographic structure etc. However, they find that only a few determinants are relevant, and that the extent to which each determinant is able to provide a credible explanation varies across countries due to the heterogeneity of their socio-economic characteristics.

[32] This information can also be very useful for policy makers and health authorities to plan the relaxation of any future national lockdown.

[33] Other coefficients are not shown for space limitations - the complete set of coefficients is available from the authors upon request.

a relatively large share of workers in the service sector. In contrast, the epidemic is weaker in provinces with a relatively large share of workers in the agriculture sector. The risk of contagion in the service sector is likely much larger than in the agricultural sector because whereas as many service jobs are indoor, tasks in the agricultural sector are mainly outdoor.

4.3.3. Relaxing distributional assumptions: NLLS results

The estimation of our frontier epidemic model comprises the parameters of the frontier epidemic curve, the parameters of the scaling function, and the structure of the two error components (i.e., the variance of v_{it} and u_i). These parameters have been estimated simultaneously in a single stage by ML once a set of (perhaps strong) distributional assumptions on both random terms have been made.

Given that our model possesses the scaling property, distributional assumptions can be relaxed somewhat. In particular, Simar et al. (1994) show that some parts of a model possessing the scaling property can be estimated using a method-of-moments (MM) approach without making distributional assumptions. Following the MM approach introduced by these authors, the parameters of both the frontier epidemic curve and the scaling function can be estimated in a first stage using a non-linear least squares (NLLS) estimator. This stage is independent of distributional assumptions with respect to the composed error component, except for it having a zero mean. Distributional assumptions are only invoked in the second stage, in which we obtain ML estimates of the parameter(s) describing the variance of v_{it}, conditional on the parameters estimated in the first stage. To see this, note that once we assume (12), our regression model can be rewritten as $\Delta lnY_{it} = ln\beta_{it}^{*}(\cdot) - \Delta h_{it}\mu + \tilde{\varepsilon}_{it}$, where $\tilde{\varepsilon}_{it} = v_{it} - \Delta h_{it} \cdot (u_i - \mu)$, and $\mu = E(u_i) \geq 0$. As $E(\tilde{\varepsilon}_{it}) = 0$, this equation can be estimated by NLLS. Given that we have already assumed that $u_i \sim N^{+}(0, \sigma_u)$, this implies that we immediately have an estimate of σ_u using the first-stage estimate of μ from the expression $\hat{\sigma}_u = \hat{\mu}\sqrt{\pi/2}$. Thus, only the parameter(s) describing the variance of v_{it} should be estimated in the second-stage of the procedure.

The NLLS parameter estimates of our linear and exponential epidemic-time and SIR-based models are provided in Tables 6 and 7 respectively. As with our previous models estimated by ML, most of our NLLS-estimated models provide evidence that the mobility of the people across provinces did clearly spread the virus across the country. The results also suggest that the lockdown has been effective in preventing the propagation of the coronavirus *between* provinces. Regarding the two random terms, we again find a decreasing standard deviation for the noise term. The parameter estimates of the scaling function vary notably, although we always find a significant coefficient for at least one of the time-varying variables of the scaling function.

[Insert Table 6 here]

[Insert Table 7 here]

The most important differences between MLE and NLLS-estimated models are found when we compute the rates of (under)reporting of coronavirus cases. These rates heavily depend on the parameter measuring the standard deviation of u_i. While this parameter is often close to 1 when using our MLE-estimated models, it rises to 2 or higher when we use NLLS techniques. We also find smaller estimates for the standard deviation of v_{it}. Consequently, the NLLS-estimated models provide extremely low reporting rates that range on average from 0% to 7%, and the multiplication factors obtained using NLLS techniques therefore seem to be seriously

upwardly-biased. Apart from differences in size, the coefficient of correlation between reporting rates from MLE and NLLS-estimated models is relatively low (close to 60%) when we use the epidemic-time specification for the epidemic curve. Interestingly, the coefficient of correlation rises to 87% when we use the SIR-based specification. This seems to suggest that the alleged bias decreases when we use a frontier specification based on a theoretical epidemiological model.

Whether this is a data-driven result or has to do with the NLLS technique itself is an empirical issue. To examine this, we carried out the same simulation exercises that were discussed in Section 3, but now using a NLLS estimator.[34] These simulations clearly indicated that the NLLS estimator tends to produce larger standard deviations of u_i than the MLE estimator when we use an epidemic-time specification for the epidemic curve. Moreover, NLLS tends to overestimate the true standard deviation of u_i, thereby seriously underestimating the true reporting rates. The performance of the NLLS estimator was even poorer when we used a SIR-based specification of the model. We find extremely poor estimates of the true proportions of undocumented cases, likely caused by lack of variations in our estimates of u_{it}.

4.3.4. Temporal windows

We also examine the effect of different temporal windows when estimating our epidemiological frontier model. As most control measures began on the days of March 13th and 14th, the data used in our empirical analysis on coronavirus cases corresponded to a temporal window defined between the onset of the epidemic in each province and the 4th of April (i.e., about three weeks before and three weeks after mid-March). The sample epidemic time ranges from $K_{it} = 3$ to $K_{it} = 40$ in this window, labelled hereafter as W0340. The first two days of the epidemic of each province are not used because we need two temporal lags to estimate the SIR-based models.

As mentioned above, zero rates of growth of coronavirus cases often appear at the beginning of outbreaks. We estimated our epidemic models dropping these observations, for two reasons. First, we found convergence problems when estimating the frontier specifications of our epidemic curves. Second, when using non-frontier econometric techniques we found that only the initial temporal patterns were biased once we dropped observations with zero rates of growth of coronavirus cases (we use a third-order function of lnK_{it}).

In order to partially address this issue of dropping observations, we re-estimate our models using two additional alternative temporal windows. The epidemic time ranges from $K_{it} = 7$ to $K_{it} = 44$ in the first window (W0744 hereafter) and ranges from $K_{it} = 10$ to $K_{it} = 47$ in the second window (W1047 hereafter). As we move from the first through to the third window, there is a fall in the number of zero rates of growth dropped from the sample. Whereas in the first (original) window we dropped 134 observations with zero rates of growth of coronavirus cases, this figure falls by half in the second window (67 observations were dropped in W0744), and falls by half again in the third and final window (only 30 observations were dropped in W1047).

While the panel datasets for each window are highly unbalanced due to the widely-differing epidemic onset dates across provinces, the second and third windows use more complete panel

[34] These simulations were not included in Section 3 because our focus there was on the specification of the frontier epidemic curve and not with the approach selected to estimate the model. The simulations are available from the authors upon request.

datasets. They do, however, reduce the number of pre-lockdown observations, which is problematic in that these are needed not only to measure the effectiveness of the Spanish lockdown to battle the COVID-19 pandemic but also to estimate spatial propagation effects across the Spanish provinces. As such, there are advantages and disadvantages to using windows that begin at later dates. To assess these trade-offs, we present the parameter estimates of the exponential epidemic-time specification for the three different temporal windows (W0340, W0744, W1047). The parameter estimates are presented in Table 8.

[Insert Table 8 here]

As the volatility of the rates of growth of reported cases is much larger in the earlier stages of the epidemic, the goodness-of-fit increases notably in the second and third windows. We find similar provincial reporting rates, with correlation coefficients close to 90% in all cases. The temporal patterns of these reporting rates are also similar, although the reporting rates are larger in the later windows. On the other hand, we do not find significant spatial propagation effects across provinces when we use the second and third windows because they include much fewer pre-lockdown observations, a result that is to be expected. As the national lockdown of the population basically halted the mobility of people across provinces, this effect can only be measured if there is a relatively large dispersion of epidemic developments across provinces before the implementation of the Spanish lockdown. Using the final window (W1047), we do not find a significant effect of the lockdown on the rates of growth of coronavirus cases. Again, this is to be expected because W1047 includes fewer of the pre-lockdown observations that are needed to identify a differential temporal pattern before and after the policy measure.

4.3.5. Variance-covariance matrix specification

As a traditional two-sided error term was simply added in order to capture measurement errors in the *rate of growth* of coronavirus cases, our noise term in equation (11) is not autocorrelated over time. This term is no longer independently distributed over time if we introduce the noise term before computing the rates of growth of coronavirus cases, in the spirit of Chudik et al. (2020) and Millimet and Parmeter (2020). We now assume that the noise term capturing measurement errors in the rate of growth of coronavirus cases is autocorrelated over time, in the same fashion as Wang and Ho (2010).

When varying the temporal windows above, we found that the number of zero rates of growth dropped from the sample decreases when we used more recent temporal windows. Moreover, we did not find severe convergence issues when we estimated W1047 using all observations because this window only includes 30 observations with zero rates of growth of coronavirus cases. In what follows, we use this temporal window to compare the parameter estimates of two models that uses the same specification of the epidemic curve (i.e., the exponential epidemic-time specification) but two different specifications for the variance-covariance matrix of our noise term. The panel datasets used in both cases do not have missing observations in between $t = 1$ and $t = T$. As the rates of growth in this window vary less than in the earlier temporal windows, we only use a second-order function of lnK_{it} to depict the epidemic curve. The results are presented in Table 9.

[Insert Table 9 here]

Generally speaking, we find that our results based on a diagonal definition of the variance-covariance matrix of the noise term are quite robust. Moreover, this specification outperforms the alternative specification as the goodness-of-fit of the model using (31) is larger than the goodness-of-fit of the model using (34). We find that the coefficients of lnK_{it} and lnK_{it}^2 are

statistically significant in both models. Again, we do not find significant spatial propagation effects across provinces; nor do we find significant effects of the lockdown on the rates of growth of coronavirus cases, due to lack of pre-lockdown observations in this window. The frontier coefficients are thus robust to different specifications of the variance-covariance matrix of the noise term. On the other hand, regardless of whether we use (31) or (34) to model the variance-covariance matrix of the noise term, we find that its standard deviation decreases with lnK_{it}. This feature of the noise term is thus robust to the specification of the noise term as autoregressive. Finally, we find that most of the coefficients of the scaling function are negative using both specifications, indicating again that the proportion of undocumented (reported) cases decreases (increases) over time. We only get a larger increase of reported cases using the autoregressive variance-covariance matrix (34).

4.4. Discussion

Despite the relatively modest (logged) standard deviations of u_i found 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 often relies on common rates. For instance, the strength of the government mitigation policy is modelled in Chudik et al. (2020) in terms of the proportion of population that is exposed to COVID-19. To estimate this proportion, they need to make an assumption regarding the reporting rate. In particular, they use the data from the Diamond Princess cruise ship reported by Moriarty et al. (2020) to calibrate this rate and assume that the average reporting rate is equal to 50% in *all* Chinese provinces. They find a very large exposure rate in Hubei province (the epicentre of the epidemic), where reducing this exposure required time due to the novelty of the virus. The estimated exposure rates in other provinces range between 9% and 87%, indicating that the Chinese control measures did have very different effects in each province.[35] This somewhat unexpected result might be caused by the common value used by these authors to calibrate the reporting rate. 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 might 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 has been effective in both preventing the propagation of the coronavirus between provinces, as well as in attenuating the propagation of the virus within each province. In other words, 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 are 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

[35] The estimated effectiveness of the social distancing policies is robust to using province-specific or pooled parameters and large or shorter temporal periods.

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 when planning the relaxation of a lockdown. We find that the most-populated provinces had more intensive coronavirus epidemics. More (less) intensive coronavirus epidemics are also found in provinces with a relatively large share of workers in the service (agricultural) sector. These results, together the strong propagation effects estimated for provinces close to the main epicentre of the coronavirus in Spain, suggest carrying out a gradual, focused relaxation of the control measures in Spain. Thus, the relaxation of the lockdown should likely 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.[36]

To conclude our discussion, it should be noted that another mitigation measure often implemented by the health authorities and government is the implementation of massive testing programs in order to uncover asymptomatic (or undocumented) coronavirus cases. Our results seem to support this type of measure because the cumulative incidence of COVID-19 tends to decrease with the initial reporting rates. We also find that there is a direct and strong relationship between the reporting rates and the onset of the epidemic. The earlier the onset day, the smaller the reporting rates.[37] Therefore, our findings suggest prioritizing the detection of coronavirus cases at early stages of the epidemics as an effective strategy to combat the propagation of this virus.

5. Conclusions and future research.

This is one of the first papers that attempts to bridge the epidemiological modelling and production economics literatures by proposing stochastic frontier analysis as a useful tool with which the epidemic curves of COVID-19 can be estimated. We have proposed two different types of stochastic epidemic frontier specifications, one based on the econometric SIR specification of Chudik et al. (2020) and the other based on previous work by Orea and Álvarez (2020) which approximates the epidemic curves with functions of the epidemic times, i.e., the time since the onset of the pandemic. The most appealing feature of these models is that they

[36] As most tasks in the construction sectors are outdoor, this sector might also be restarted before other sectors.

[37] This result simply provides evidence about the difficulty to detect this new and little-known virus by the Spanish regional health systems.

both can be estimated using standard stochastic frontier techniques. One of the specifications of the model can be interpreted as a heteroskedastic version of the model introduced by Wang and Ho (2010). As such, the model we propose should prove useful for practitioners to control for individual effects in a production economics context under time-varying heteroskedasticity.

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 wave 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. One way to capture this cross-group propagation effect would be to use a two-step procedure where, in the first step, a standard DEA is used to obtain an estimate of the proportion of undocumented cases and then, in the second step, this estimate is included as a regressor in the epidemic model of reported cases. The estimated coefficient of this variable shows the elasticity of reported cases with respect to changes in the proportion (number) of undocumented cases, and therefore can be used to test the so-called cross-group propagation effect. Adetutu et al. (2016) adopted a similar two-stage strategy to produce wide range of rebound effects from super-conservation to backfire.

References

Adetutu, M.O., Glass, A.J., Weyman-Jones, T.G., 2016. Economy-wide estimates of rebound effects: Evidence from panel data. Energy Journal 37(3), 251-269. http://www.jstor.org/stable/44075659

Battese, G., Coelli, T., 1992. Frontier production functions, technical efficiency and panel data: with application to paddy farmers in India. Journal of Productivity Analysis 3, 153-169. https://doi.org/10.1007/BF00158774.

Chudik, A., Pesaran, M. H., Rebucci, A., 2020. Voluntary and mandatory social distancing: Evidence on Covid-19 exposure rates from Chinese provinces and selected countries. NBER Working paper 27039. Working Paper 27039, http://www.nber.org/papers/w27039.

Eliasson, K., Lindgren, U., Westerlund, O., 2003. Geographical labour mobility: migration or commuting? Regional studies 37(8), 827-837. https://doi.org/10.1080/0034340032000128749

Flaxman, S., Mishra, S., Gandy, A., Unwin, H.J.T., Mellan, T.A., Coupland, H., Whittaker, C., Zhu, H., Berah, T., Eaton, J.W., Monod, M., 2020. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. Nature 584(7820), 257-261. https://www.nature.com/articles/s41586-020-2405-7.

Giuliani, D., Dickson, M.M., Espa, G., Santi, F., 2020. Modelling and predicting the spatio-temporal spread of Coronavirus disease 2019 (COVID-19) in Italy. Available at SSRN: https://ssrn.com/abstract=3559569 or http://dx.doi.org/10.2139/ssrn.3559569.

Google, 2020. COVID-19 Community Mobility Report. Spain April 11, 2020. https://www.gstatic.com/covid19/mobility/2020-04-11_ES_Mobility_Report_en.pdf.

Greene, W., 2005. Reconsidering heterogeneity in panel data estimators of the stochastic frontier model. Journal of Econometrics 126 (2), 269-303. https://doi.org/10.1016/j.jeconom.2004.05.003

Gross, B., Zheng, Z., Liu, S., Chen, X., Sela, A., Li, J., Li, D., Havlin, S., 2020. Spatio-temporal propagation of COVID-19 pandemics. Available at medRxiv preprint doi: https://doi.org/10.1101/2020.03.23.20041517.

Harris, R.D.F., Tzavalis, E., 1999. Inference for unit roots in dynamic panels where the time dimension is fixed. Journal of Econometrics 91(2), 201-226. https://doi.org/10.1016/S0304-4076(98)00076-1

Korolev, I., 2021. Identification and Estimation of the SEIRD Epidemic Model for COVID-19. Journal of Econometrics 220(1), 63-85. https://doi.org/10.1016/j.jeconom.2020.07.038

Kumbhakar, S.C., 1990. Production frontiers, panel data, and time-varying technical inefficiency. Journal of Econometrics 46, 201–211. https://doi.org/10.1016/0304-4076(90)90055-X

Lai, H-P., Huang, C.J., 2013. Maximum likelihood estimation of seemingly unrelated stochastic frontier regressions. Journal of Productivity Analysis 40(1), 1-14. https://doi.org/10.1007/s11123-012-0289-8

Leung, K., Wu, J. T., Liu, D., Leung, G. M., 2020. First-wave COVID-19 transmissibility and severity in China outside Hubei after control measures, and second-wave scenario planning: a modelling impact assessment. The Lancet 395, 1382-1393. https://doi.org/10.1016/S0140-6736(20)30746-7.

Li, R., Pei, S., Chen, B., Song, Y., Zhang, T., Yang, W., Shaman, J., 2020. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV2). Science 368 (6490), 489-493. https://doi.org/10.1126/science.abb3221

Millimet, D.L., Parmeter, C.F., 2020. COVID-19 Severity: A New Approach to Quantifying Global Cases and Deaths, University of Miami, unpublished document.

Moriarty, L., M. Plucinski, B. Marston, et al., 2020. Public health responses to COVID-19 outbreaks on cruise ships - worldwide, February-March 2020. Morbidity and Mortality Weekly Report (MMWR), 26 March 2020, 69:347-352. https://doi.org/10.15585/mmwr.mm6912e3.

Orea L., Alvarez, I., 2020. How effective has the Spanish lockdown been to battle COVID-19? A spatial analysis of the coronavirus propagation across provinces. Working Paper 2020/3, FEDEA (Fundación de Estudios de Economía Aplicada). Available at: https://documentos.fedea.net/pubs/dt/2020/dt2020-03.pdf

Orea, L., Álvarez, I.C., 2019. A new stochastic frontier model with cross-sectional effects in both noise and inefficiency terms. Journal of Econometrics 213(2), 556-577. https://doi.org/10.1016/j.jeconom.2019.07.004

Simar, L., Lovell, C.A.K., Vanden Eeckaut, P., 1994. Stochastic frontiers incorporating exogenous influences on efficiency. Discussion paper No. 9403, Institute de Statistique, UCL-Université Catholique de Louvain.

Stojkoski, V., Utkovski, Z., Jolakoski, P., Tevdovski, D., Kocarev, L., 2020. The socio-economic determinants of the coronavirus disease (COVID-19) pandemic. arXiv preprint arXiv:2004.07947.

Vega, S.H., Elhorst, J.P., 2015. The SLX model. Journal of Regional Science 55(3), 339–363. https://doi.org/10.1111/jors.12188

Wang, H.-J., 2003. A stochastic frontier analysis of financing constraints on investment: The case of financial liberalization in Taiwan. Journal of Business and Economic Statistics 21, 406-419. https://doi.org/10.1198/073500103288619016.

Wang, H.J., Ho, C.W., 2010. Estimating fixed-effect panel stochastic frontier models by model transformation. Journal of Econometrics 157(2), 286-296. https://doi.org/10.1016/j.jeconom.2009.12.006

Table 1. Simulation results for sample structure ($N = 20,\ T = 50$)

Sim. No.	Parameter Settings and MF					Model 1: Epidemic time ($\ln K_t$) specification						Model 2: SIR specification					
	Parameter Values				True MF	Results						Results					
	$ln\sigma_v$	η_u	η_v	$ln\sigma_u$	e^{-u}	R^2 (u	R^2 (u	$ln\sigma_u$	MSE (u	Corr.$(u,\hat{u})$	$\hat{\eta}_u$	R^2 (u	R^2 (u	$ln\sigma_u$	MSE (u	Corr.$(u,\hat{u})$	$\hat{\eta}_u$
1a	-3	-	0	1	1.289 *(0.139)*	0.469 *(0.048)*	0.635 *(0.057)*	0.951 *(0.185)*	0.001 *(0.002)*	0.997 *(0.002)*	-0.203 *(0.007)*	0.618 *(0.055)*	0.727 *(0.043)*	1.605 *(0.166)*	0.061 *(0.026)*	0.997 *(0.001)*	-0.226 *(0.005)*
1b	-3	-	0	0.5	1.121 *(0.026)*	0.384 *(0.041)*	0.475 *(0.050)*	0.413 *(0.208)*	0.002 *(0.002)*	0.993 *(0.004)*	-0.207 *(0.012)*	0.485 *(0.037)*	0.608 *(0.035)*	1.118 *(0.161)*	0.021 *(0.011)*	0.993 *(0.003)*	-0.232 *(0.011)*
2a	-3	-	-	1	1.269 *(0.088)*	0.716 *(0.055)*	0.961 *(0.011)*	1.547 *(0.153)*	0.114 *(0.055)*	0.950 *(0.021)*	-0.198 *(0.005)*	0.919 *(0.021)*	0.966 *(0.009)*	1.633 *(0.143)*	0.089 *(0.027)*	0.995 *(0.002)*	-0.216 *(0.004)*
2b	-3	-	0.1	0.5	1.120 *(0.029)*	0.743 *(0.041)*	0.929 *(0.016)*	1.251 *(0.264)*	0.098 *(0.066)*	0.931 *(0.032)*	-0.199 *(0.009)*	0.855 *(0.029)*	0.939 *(0.013)*	1.265 *(0.132)*	0.060 *(0.019)*	0.989 *(0.005)*	-0.210 *(0.007)*
3a	-3	-	0	1	2.015 *(0.429)*	0.422 *(0.039)*	0.565 *(0.049)*	0.892 *(0.178)*	0.011 *(0.013)*	0.996 *(0.003)*	-0.106 *(0.005)*	0.598 *(0.041)*	0.672 *(0.037)*	1.402 *(0.221)*	0.053 *(0.041)*	0.988 *(0.003)*	-0.135 *(0.005)*
3b	-3	-	0	0.5	1.346 *(0.082)*	0.335 *(0.035)*	0.406 *(0.046)*	0.318 *(0.198)*	0.022 *(0.109)*	0.975 *(0.155)*	-0.111 *(0.015)*	0.479 *(0.031)*	0.551 *(0.032)*	0.825 *(0.212)*	0.012 *(0.006)*	0.979 *(0.009)*	-0.148 *(0.013)*
4a	-2	-	0	1	1.285 *(0.101)*	0.146 *(0.029)*	0.196 *(0.039)*	0.788 *(0.259)*	0.016 *(0.010)*	0.980 *(0.011)*	-0.226 *(0.030)*	0.373 *(0.036)*	0.400 *(0.036)*	0.575 *(1.237)*	0.042 *(0.054)*	0.904 *(0.200)*	-0.439 *(0.447)*
4b	-2	-	0	0.5	1.118 *(0.026)*	0.088 *(0.016)*	0.1 *(0.024)*	0.032 *(0.670)*	0.186 *(0.308)*	0.634 *(0.558)*	-0.205 *(0.330)*	0.314 *(0.023)*	0.336 *(0.027)*	0.401 *(0.721)*	0.021 *(0.014)*	0.888 *(0.105)*	-0.477 *(0.363)*

Table 2. Simulation results for sample structure ($N = 40,\ T = 25$)

Sim. No.	Parameter Settings and MF					Model 1: Epidemic time ($\ln K_t$) specification						Model 2: SIR specification					
	Parameter Values				True MF	Results						Results					
	$ln\sigma_v$	η_u	η_v	$ln\sigma_u$	e^{-u}	R^2 $(u=0)$	R^2 $(u\geq 0)$	$ln\sigma_u$	MSE $(u-\hat{u})$	Corr. $(u,\hat{u})$	$\hat{\eta}_u$	R^2 $(u=0)$	R^2 $(u$	$ln\sigma_u$	MSE $(u-\hat{u})$	Corr. $(u,\hat{u})$	$\hat{\eta}_u$
1a	-3	-0.2	0	1	1.562 *(0.134)*	0.436 *(0.041)*	0.704 *(0.042)*	0.931 *(0.134)*	0.003 *(0.002)*	0.997 *(0.002)*	-0.202 *(0.007)*	0.728 *(0.031)*	0.777 *(0.032*	0.622 *(0.218)*	0.042 *(0.014)*	0.944 *(0.027)*	-0.108 *(0.018)*
1b	-3	-0.2	0	0.5	1.234 *(0.044)*	0.332 *(0.038)*	0.503 *(0.055)*	0.393 *(0.150)*	0.003 *(0.002)*	0.993 *(0.003)*	-0.209 *(0.010)*	0.571 *(0.042)*	0.622 *(0.042*	0.131 *(0.176)*	0.018 *(0.009)*	0.937 *(0.030)*	-0.106 *(0.020)*
2a	-3	-0.2	-0.1	1	1.566 *(0.146)*	0.581 *(0.041)*	0.948 *(0.012)*	0.970 *(0.123)*	0.001 *(0.001)*	0.998 *(0.001)*	-0.203 *(0.005)*	0.918 *(0.014)*	0.962 *(0.009*	1.141 *(0.182)*	0.032 *(0.020)*	0.997 *(0.002)*	-0.178 *(0.008)*
2b	-3	-0.2	0.1	0.5	1.229 *(0.033)*	0.588 *(0.039)*	0.885 *(0.018)*	0.458 *(0.108)*	0.001 *(0.001)*	0.996 *(0.002)*	-0.206 *(0.009)*	0.845 *(0.021)*	0.915 *(0.014*	0.610 *(0.140)*	0.007 *(0.004)*	0.996 *(0.002)*	-0.186 *(0.008)*
3a	-3	-0.1	0	1	3.049 *(0.883)*	0.293 *(0.034)*	0.577 *(0.048)*	0.902 *(0.121)*	0.013 *(0.013)*	0.997 *(0.002)*	-0.101 *(0.003)*	0.612 *(0.033)*	0.683 *(0.035*	1.275 *(0.249)*	5.400 *(9.867)*	0.815 *(0.065)*	-0.023 *(0.008)*
3b	-3	-0.1	0	0.5	1.688 *(0.150)*	0.219 *(0.026)*	0.379 *(0.047)*	0.383 *(0.150)*	0.011 *(0.007)*	0.992 *(0.003)*	-0.102 *(0.005)*	0.478 *(0.033)*	0.536 *(0.039*	1.258 *(0.437)*	11.827 *(22.727)*	0.760 *(0.050)*	-0.013 *(0.006)*
4a	-2	-0.2	0	1	1.573 *(0.144)*	0.153 *(0.026)*	0.249 *(0.040)*	0.812 *(0.191)*	0.031 *(0.019)*	0.979 *(0.010)*	-0.227 *(0.028)*	0.441 *(0.030)*	0.441 *(0.030*	0.814 *(0.224)*	2.798 *(0.897)*	-0.061 *(0.568)*	0.000 *(0.000)*
4b	-2	-0.2	0	0.5	1.233 *(0.040)*	0.080 *(0.015)*	0.107 *(0.028)*	-0.472 *(1.642)*	0.055 *(0.040)*	0.801 *(0.334)*	-0.379 *(0.342)*	0.346 *(0.022)*	0.346 *(0.022*	0.362 *(0.167)*	1.115 *(0.366)*	-0.075 *(0.657)*	0.000 *(0.000)*

Table 3. MLE: Epidemic-time (lnK_{it}) specification

	Linear				Exponential			
	Non-frontier model		Frontier model		Non-frontier model		Frontier model	
	Coef.	s.e.	Coef.	s.e.	Coef.	s.e.	Coef.	s.e.
Overall epidemic curve								
$Intercept$	0.8083 **	0.3734	0.8950	0.7857	4.1081 ***	1.1877	5.1474	3.2768
lnK_{it}	-0.3775	0.4130	-0.4959	0.8168	-7.0363 ***	1.5375	-8.8926 **	3.8386
lnK_{it}^2	0.1492	0.1518	0.1602	0.2869	3.2610 ***	0.6438	4.0997 ***	1.5123
lnK_{it}^3	-0.0248	0.0183	-0.0227	0.0333	-0.5010 ***	0.0863	-0.6332 ***	0.1964
$W_i lnK_{it}$	0.0360 *	0.0185	0.0830 ***	0.0201	0.0835 ***	0.0499	0.1278 *	0.0724
D_t	-0.1815 ***	0.0304	-0.1376 ***	0.0295	-0.5964 ***	0.0875	-0.4977 ***	0.1306
$W_i lnK_t \cdot D_t$	-0.0466 ***	0.0187	-0.0782 ***	0.0175	-0.1964 ***	0.0541	-0.2879 ***	0.0592
Noise term ($ln\sigma_v$)								
$Intercept$	0.7954 ***	0.1375	1.1092 ***	0.0924	0.8430 ***	0.1333	1.1247 ***	0.1111
lnK_{it}	-1.0215 ***	0.0485	-1.1673 ***	0.0322	-1.0411 ***	0.0468	-1.1714 ***	0.0390
Scaling function								
K_t			-0.0383 ***	0.0112			-0.0437 ***	0.0145
$W_i lnK_t$			-0.1376 ***	0.0429			-0.0796	0.0501
$K_t \cdot D_t$			-0.0020 *	0.0011			-0.0026 *	0.0014
$W_i lnK_t \cdot D_t$			-0.0281 **	0.0136			-0.0350 *	0.0194
u-term ($ln\sigma_u$)								
$Intercept$			1.2122 ***	0.5060			1.1527 ***	0.4280
Day of the week effects	Yes		Yes		Yes		Yes	
Mean log LF	0.6572		1.6373		0.6648		1.6379	
Pseudo R-sq	0.3442		0.3735		0.3291		0.3507	
Mean RR			0.3490				0.4220	
Obs.	1290		1290		1290		1290	

Notes: *, **, *** indicate significance at the 10, 5, 1% level, respectively.

Table 4. MLE: SIR specification

	Linear				Exponential			
	Non-frontier model		Frontier model		Non-frontier model		Frontier model	
	Coef.	s.e.	Coef.	s.e.	Coef.	s.e.	Coef.	s.e.
Overall epidemic curve								
$Intercept$	0.2897 ***	0.0301	0.1842 ***	0.0385	-1.4418 ***	0.1058	-2.0487 ***	0.1978
lnY_{t-1}	0.1948 ***	0.0230	0.0370	0.0369	0.2666 ***	0.0679	-0.0901	0.1467
lnY_{t-2}	-0.2380 ***	0.0220	-0.1062 ***	0.0330	-0.5327 ***	0.0618	-0.3994 ***	0.1238
$lnY_{t-1} \cdot lnY_{t-2}$	0.0053 ***	0.0007	0.0020	0.0020	-0.0105 **	0.0053	-0.0334 **	0.0149
$W_i lnK_t$	0.0550 ***	0.0181	0.0972 ***	0.0218	0.1941 ***	0.0491	0.4492 ***	0.1442
D_t	-0.1121 ***	0.0296	-0.0574	0.0363	-0.3384 ***	0.0875	-0.0790	0.1639
$W_i lnK_{it} \cdot D_t$	-0.0591 ***	0.0182	-0.0863 ***	0.0187	-0.2344 ***	0.0524	-0.3795 ***	0.1157
Noise term $(ln\sigma_v)$								
$Intercept$	0.8743 ***	0.1352	1.0180 ***	0.0926	0.6446 ***	0.1288	0.9607 ***	0.1537
lnK_t	-1.0710 ***	0.0476	-1.1553 ***	0.0326	-0.9807 ***	0.0453	-1.1295 ***	0.0554
Scaling function								
K_t			-0.0091	0.0233			-0.0538 ***	0.0188
$W_i lnK_{it}$			-0.0185	0.0395			-0.0688 *	0.0370
$K_t \cdot D_t$			0.0005	0.0012			0.0002	0.0018
$W_i lnK_t \cdot D_t$			0.0044	0.0103			0.0042	0.0284
u-term $(ln\sigma_u)$								
$Intercept$			2.5189	2.1028			1.2793 ***	0.2587
Day of the week effects	Yes		Yes		Yes		Yes	
Mean log LF	0.7174		1.6843		0.6934		1.6817	
Pseudo R-sq	0.3799		0.4531		0.3827		0.4344	
Mean RR			0.0340				0.4060	
Obs	1290		1290		1290		1290	

Notes: *, **, *** indicate significance at the 10, 5, 1% level, respectively.

Table 5. MLE: Exponential epidemic-time (lnK_{it}) specification with socio-economic determinants

	Coef.		s.e.	t-stat	LR test	Mean log LF
Overall epidemic curve						
$Age\ 15-25$ (%)	-0.0065		0.0456	-0.14	0.03	1.63789
$Age\ 25-65$ (%)	0.0131		0.0407	0.32	0.10	1.63792
$Age > 65$ (%)	0.0010		0.0115	0.09	0.00	1.63788
$Service$ (%)	0.0256	***	0.0090	2.84	10.32	1.64188
$Agriculture$ (%)	-0.0335	*	0.0173	-1.94	9.88	1.64171
$lnDensity$	0.1269	*	0.0703	1.81	6.40	1.64036
$lnPopulation$	0.1929	**	0.0926	2.08	11.15	1.6422
$lnGDP_{pc}$	0.1498		0.2323	0.64	0.41	1.63804
$Temperature$	0.0138		0.0141	0.98	1.75	1.63856
$Rainfall$	-0.0005		0.0061	-0.08	0.03	1.63789
Scaling function						
$Age\ 15-25$ (%)	0.0727		0.2074	0.35	0.34	1.63801
$Age\ 25-65$ (%)	0.1509		0.1975	0.76	1.50	1.63846
$Age > 65$ (%)	-0.0335		0.0643	-0.52	0.85	1.63821
$Service$ (%)	0.0204		0.0498	0.41	0.83	1.6382
$Agriculture$ (%)	-0.0232		0.0481	-0.48	1.16	1.63833
$lnDensity$	0.0738		0.2663	0.28	0.39	1.63803
$lnPopulation$	0.1070		0.3139	0.34	0.67	1.63814
$lnGDP_{pc}$	0.2266		1.1090	0.20	0.10	1.63792
$Temperature$	0.0030		0.0642	0.05	0.00	1.63788
$Rainfall$	0.0213		0.1184	0.18	0.15	1.63794

Notes:

- Each variable has been introduced one at a time.
- Other coefficients are not shown for space limitations. All coefficients are available from the authors upon request.
- *, **, *** indicate significance at the 10, 5, 1% level, respectively.
-

Table 6. NLLS: Epidemic-time (lnK_{it}) specification

	Linear				Exponential			
	Non-frontier model		Frontier model		Non-frontier model		Frontier model	
	Coef.	s.e.	Coef.	s.e.	Coef.	s.e.	Coef.	s.e.
Overall epidemic curve								
$Intercept$	0.9862 ***	0.2756	0.5872 *	0.3545	0.9910	0.8877	4.7994	19.8171
lnK_{it}	-0.4893	0.3383	0.2884	0.5696	-2.0803	1.2809	-8.5911	32.0505
lnK_{it}^2	0.1717	0.1386	-0.2373	0.2680	0.9778 *	0.5932	3.8408	14.6131
lnK_{it}^3	-0.0255	0.0182	0.0472	0.0426	-0.1748 **	0.0857	-0.5638	1.9899
$W_i lnK_t$	0.0629 ***	0.0228	0.0302	0.0231	0.1687 ***	0.0545	0.4453	0.4421
D_t	-0.2079 ***	0.0411	-0.1664 ***	0.0429	-0.6176 ***	0.0992	-2.1014	4.9969
$W_i lnK_{it} \cdot D_t$	-0.0804 ***	0.0231	-0.0844 ***	0.0228	-0.2966 ***	0.0592	-1.0258	2.3479
Noise term ($ln\sigma_v$)								
$Intercept$			0.8916 ***	0.1351			0.9026 ***	0.1326
lnK_t			-1.0756 ***	0.0478			-1.0802 ***	0.0467
Scaling function								
K_t			0.0401 ***	0.0056			-0.0388	0.0263
$W_i lnK_t$			-0.0900 **	0.0453			-0.0693 **	0.0340
$K_t \cdot D_t$			-0.0001	0.0008			-0.0012	0.0014
$W_i lnK_t \cdot D_t$			0.0023	0.0064			0.0004	0.0077
u-term ($ln\sigma_u$)								
$Intercept$			1.2378 *	0.6543			1.9489 ***	0.4810
Day of the week effects	Yes		Yes		Yes		Yes	
Mean log LF								
Pseudo R-sq	0.3498		0.3753		0.3523		0.3676	
Mean RR			0.0020				0.0710	
Obs	1290		1290		1290		1290	

Notes: *, **, *** indicate significance at the 10, 5, 1% level, respectively.

Table 7. NLLS: SIR specification

	Linear				Exponential			
	Non-frontier model		Frontier model		Non-frontier model		Frontier model	
	Coef.	s.e.	Coef.	s.e.	Coef.	s.e.	Coef.	s.e.
Overall epidemic curve								
$Intercept$	0.3340 ***	0.0382	0.1553	0.1018	-1.2951 ***	0.1208	-2.4057 ***	0.6045
lnY_{t-1}	0.0524 *	0.0320	0.0419	0.0325	0.0111	0.0901	-0.1029	0.2126
lnY_{t-2}	-0.0984 ***	0.0307	-0.0783 ***	0.0323	-0.2659 ***	0.0799	-0.4472 *	0.2324
$lnY_{t-1} \cdot lnY_{t-2}$	0.0050 ***	0.0011	0.0049 ***	0.0011	-0.0088	0.0066	-0.0291	0.0262
$W_i lnK_t$	0.0837 ***	0.0224	0.0734 ***	0.0226	0.2325 ***	0.0567	0.4488 **	0.1927
D_t	-0.1573 ***	0.0364	-0.1237 ***	0.0395	-0.4280 ***	0.0992	-0.5818 **	0.2806
$W_i lnK_t \cdot D_t$	-0.0924 ***	0.0224	-0.0813 ***	0.0224	-0.2758 ***	0.0612	-0.4960 **	0.2036
Noise term ($ln\sigma_v$)								
$Intercept$			0.7563 ***	0.1331			0.6707 ***	0.1311
lnK_t			-1.0374 ***	0.0469			-1.0041 ***	0.0462
Scaling function								
K_t			-0.0126 **	0.0053			-0.0238 ***	0.0067
$W_i lnK_t$			-0.0288	0.0267			-0.0627 ***	0.0238
$K_t \cdot D_t$			0.0000	0.0004			-0.0005	0.0009
$W_i lnK_t \cdot D_t$			0.0018	0.0037			0.0042	0.0070
u-term ($ln\sigma_u$)								
$Intercept$			2.5197 ***	0.8417			1.9251 ***	0.1856
Day of the week effects	Yes		Yes		Yes		Yes	
Mean log LF								
Pseudo R-sq	0.4011		0.4114		0.4038		0.4205	
Mean RR			0.0000				0.0300	
Obs	1290		1290		1290		1290	

Notes: *, **, *** indicate significance at the 10, 5, 1% level, respectively.

Table 8. MLE: Epidemic-time (lnK_{it}) specification with different temporal windows

	W0340		W0744		W1047	
	Coef.	s.e.	Coef.	s.e.	Coef.	s.e.
Overall epidemic curve						
$Intercept$	5.1474	3.2768	31.3080	20.0320	2.5563	41.0100
lnK_{it}	-8.8926 **	3.8386	-39.1260 *	21.3360	-10.8850	41.3300
lnK_{it}^2	4.0997 ***	1.5123	15.4096 **	7.5498	6.1740	13.8290
lnK_{it}^3	-0.6332 ***	0.1964	-2.0246 **	0.8826	-1.0255	1.5368
$W_i lnK_t$	0.1278 *	0.0724	0.107	0.1116	-0.0035	0.1765
D_t	-0.4977 ***	0.1306	-0.4942 **	0.2522	-0.4194	0.4307
$W_i lnK_t \cdot D_t$	-0.2879 ***	0.0592	-0.3453 ***	0.1054	-0.2671	0.1872
Noise term ($ln\sigma_v$)						
$Intercept$	1.1247 ***	0.1111	2.9305 ***	0.1539	4.1323 ***	0.2357
lnK_t	-1.1714 ***	0.0390	-1.7956 ***	0.0474	-2.1826 ***	0.0707
Scaling function						
K_t	-0.0437 ***	0.0145	-0.0655 ***	0.0079	-0.0674 ***	0.0061
$W_i lnK_t$	-0.0796	0.0501	0.0119	0.0433	0.0216	0.0484
$K_t \cdot D_t$	-0.0026 *	0.0014	-0.0021 *	0.0012	-0.002	0.0015
$W_i lnK_t \cdot D_t$	-0.0350 *	0.0194	-0.0298	0.0211	-0.0286	0.0248
u-term ($ln\sigma_u$)						
Intercept	1.1527 ***	0.4280	1.7139 ***	0.4982	1.7959 ***	0.5377
Day of the week effects	Yes		Yes		Yes	
Mean log LF	1.6379		1.9594		2.2137	
Obs	1290		1357		1394	
Pseudo R-sq	0.3507		0.4157		0.4635	
Epidemic time						
Minimum	3		7		10	
Maximum	40		44		47	
Mean RR	0.422		0.438		0.473	
Zero rates of growth (#)	No		No		No	

Notes: *, **, *** indicate significance at the 10, 5, 1% level, respectively.

Table 9. Diagonal vs. first-differences variance-covariance matrix of the noise term

	W1047A		W1047B		W1047C	
	Coef.	s.e.	Coef.	s.e.	Coef.	s.e.
Overall epidemic curve						
$Intercept$	-24.5151 ***	5.9564	-29.0482 ***	6.1906	-38.8470 ***	6.0106
lnK_{it}	16.4949 ***	3.9424	19.7441 ***	4.0948	25.4290 ***	4.4757
lnK_{it}^2	-3.0285 ***	0.6332	-3.5628 ***	0.6556	-4.4960 ***	0.7296
lnK_{it}^3	-0.0299	0.1813	0.2430	0.1629	-0.0210	0.5540
$W_i lnK_t$	-0.3842	0.4481	-0.6655 *	0.3471	0.2807	0.9476
D_t	-0.2528	0.1961	-0.4942 ***	0.1612	-0.1728	0.5536
$W_i lnK_t \cdot D_t$						
Noise term ($ln\sigma_v$)	4.1414 ***	0.2183	4.3246 ***	0.2278	5.8213 ***	0.1642
$Intercept$	-2.1849 ***	0.0661	-2.2389 ***	0.0698	-2.5586 ***	0.0543
lnK_t						
Scaling function	-0.0711 ***	0.0062	-0.0634 ***	0.0058	-0.0793 ***	0.0034
K_t	0.0348	0.0484	0.0471	0.0462	0.1456 ***	0.0448
$W_i lnK_t$	-0.002	0.0015	-0.0035 **	0.0015	-0.0059	0.0041
$K_t \cdot D_t$	-0.0296	0.0251	-0.0441 *	0.0242	-0.0407	0.0325
$W_i lnK_t \cdot D_t$						
u-term ($ln\sigma_u$)	1.8568 ***	0.4922	1.7275 ***	0.5083	2.2081 ***	0.2993
Intercept	Yes		Yes		Yes	
Day of the week effects	Yes		Yes		Yes	
Mean log LF	2.2119		2.1735		1.6378	
Obs	1394		1424		1424	
Pseudo R-sq	0.4575		0.3457		0.2877	
Zero rates of growth (#)	No		Yes		Yes	
Π matrix	Diagonal		Diagonal		First-Differences	
Epidemic time						
Minimum	10		10		10	
Maximum	47		47		47	
Mean RR	0.482		0.470		0.473	

Notes: *, **, *** indicate significance at the 10, 5, 1% level, respectively.