Centre for Efficiency and Productivity Analysis

Stochastic Frontier Analysis for Healthcare, with Illustrations in R

Robin C. Sickles, Zhichao Wang, Valentin Zelenyuk

Date: May 2022

Stochastic Frontier Analysis for Healthcare, with Illustrations in R

Robin C. Sickles[*]   Zhichao Wang[†]   Valentin Zelenyuk[‡]

May 26, 2022

Abstract

In this chapter, we provide a brief overview of the stochastic frontier analysis (SFA) in the context of analysing healthcare, with a focus on hospitals, where it has received most attention. We start with the classical SFA model of Aigner, Lovell and Schmidt (1977) and then consider many of its popular extensions and generalizations in both cross-sectional and panel data (mainly published in Journal of Econometrics, Journal of Business & Economic Statistics and Journal of Productivity Analysis). We also briefly discuss semi-parametric and non-parametric generalizations, spatial frontiers, and Bayesian SFA. Whenever possible, we refer the readers to various applications of these general methods to healthcare, and for hospitals in particular. Finally, we also illustrate some of these methods for real data on public hospitals in Queensland, Australia, as well as provide practical guidance and references for their computational implementations via R.

Keywords: Stochastic frontier analysis, R, healthcare, hospital, Queensland

[*]Department of Economics, Rice University, Houston, TX 77251-1892, USA

[†]School of Economics, University of Queensland, Brisbane, Qld 4072, Australia

[‡]School of Economics and Centre for Efficiency and Productivity Analysis, University of Queensland, Brisbane, Qld 4072, Australia

1   Introduction

Stochastic frontier analysis (SFA) has become one of the most important methods of efficiency measurement since it was introduced by Aigner et al. (1977). With the many stochastic frontier models (SFMs) developed, SFA has been widely applied in various fields of research in the last few decades. This popularity is also vivid in most reviews of efficiency analysis, where SFA and Data Envelopment Analysis (DEA) are usually referred to as the two main paradigms, in general and for healthcare in particular, e.g., Lampe and Hilgers (2015) for general studies, as well as Hollingsworth (2008); See et al. (2021); Rosko (2022) for a specific field (i.e., healthcare).

In SFMs, the (in)efficiency of production (or cost, profit, etc.) of units is evaluated according to the estimated frontier. However, the approaches of the frontier and (in)efficiency estimation in different model specifications differ from each other. For a more comprehensive discussion, the readers can refer to Sickles and Zelenyuk (2019, Chapter 11-16) and Kumbhakar et al. (2022a,b) for details. In this chapter, we aim to provide a brief overview of the wide range of SFMs from the basic to the recently developed ones. Following the development route, we present representative models for different scenarios and their demonstrations in R (R Core Team, 2022). This chapter is also an update of the SFA part of Sickles et al. (2020) and Nguyen et al. (2022), calibrated and fine-tuned to the context of healthcare.

In the following chapter, we start from the basic form, on which, most models were developed. We then turn to two main streams of further developments, i.e., pursuing the advancements in solving panel data scenarios and accounting for environmental determinants of (in)efficiency. We also briefly introduce the recent development of SFMs in semi- and non-parametric generalizations, as well as the spatial and Bayesian SFMs in the context of healthcare. Alongside this, we provide demonstrations of the representative SFMs with an example of public hospitals in Queensland, Australia.

2 Basic SFM

The key feature of SFM is its composed error term of statistical noise and an additional non-negative error term.[1] The randomness and measurement error are accounted for in the ordinary noise term, while the additional error term accounts for the technical inefficiency, which is the reduced output with respect to the frontier output (therefore non-negative) (Sickles and Zelenyuk, 2019). The advancement in distinguishing the inefficiency (as a discrete error term) from the noise is usually at the cost of parametric assumptions. Yet, the more interpretable estimators of both the production function and inefficiency (and potentially the determinant factors) may be worth the constraints and computational difficulty, according to many arguments in the literature. Alternatively, non-parametric methods (e.g., DEA) can be considered to eliminate the parametric assumptions, or instead, the semi-parametric approaches may be applied to avoid the dilemma, which we will discuss later in this chapter.

Specifically, in an example context of the Cobb-Douglas production function, the basic SFM can be formulated as

$$Y_i = \beta_0 \prod_{j=1}^{p} \beta_j X_{j,i} \exp(v_i) \exp(-u_i), \quad i = 1, \dots, n, \tag{1}$$

where $X \in \mathfrak{R}_+^p$ is a vector of input variables and $Y \in \mathfrak{R}_+$ is the corresponding output. The vector of parameters to be estimated is represented by β, while $\exp(v_i)$ and $\exp(-u_i)$ stand for the statistical noise and technical efficiency, respectively. Equation (1) can be rewritten in a linear form via logarithmic transformation as

$$y_i = a + \sum_{j=1}^{p} \beta_j x_{j,i} + v_i - u_i, \quad i = 1, \dots, n, \tag{2}$$

in which $x_{j,i} = \ln(X_{j,i})$, $y_i = \ln(Y_i)$, $a = \ln(\beta_0)$ and so u_i represents the technical inefficiency, the gap between the maximum potential output and the actual output in logarithm format. Accordingly,

---

[1]The discussion hereafter is based on the (in)efficiency of production function.

different SFMs propose different assumptions on the production function and the error terms, as well as different techniques for their estimation, which we will cover in more detail in the following sections.

2.1 Aigner et al. (1977)

Aigner et al. (1977) (hereafter ALS77) and Meeusen and van den Broeck (1977) independently proposed the canonical SFM at about the same time. The ALS77 model can be formulated as

$$y_i = a + \sum_{j=1}^{p} \beta_j x_{j,i} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{3}$$

where the composed error term ε_i can be derived as

$$\varepsilon_i = v_i - u_i \quad i = 1, \ldots, n. \tag{4}$$

The random error term v_i and the inefficiency term u_i in the ALS77 model are assumed to be independently and identically distributed (*iid*) with normal and half-normal distribution, respectively, i.e.,

$$\begin{aligned} v_i &\sim iidN(0, \sigma_v^2), \\ u_i &\sim iidN^+(0, \sigma_u^2). \end{aligned} \tag{5}$$

Therefore, the ALS77, and many other variants of SFM assuming the distributions of error terms, can be estimated with the maximum likelihood estimator (MLE) using the likelihood function. After obtaining the estimated parameters, the expected level of inefficiency ($E(u)$) can be further estimated as $E(u) = \sqrt{2/\pi}\sigma_u$, and the level of efficiency ($E[exp(-u)]$) can be approximately derived as $E[exp(-u)] \approx 1 - E(u)$.

Furthermore, the level of (in)efficiency of the individual unit can be estimated with the estimators of error terms, e.g., following the most popular practice in the literature introduced by Jondrow et al. (1982) (hereafter JLMS), where the expected inefficiency term (u_i) is estimated conditionally on the composed error (ε_i). To be more concrete, the inefficiency can be estimated as

$$\widehat{u}_i = E(u_i | \varepsilon_i)$$

$$= u_{i*} + \sigma_* \frac{\phi(-u_{i*}/\sigma_*)}{1 - \Phi(-u_{i*}/\sigma_*)}, \tag{6}$$

where

$$\sigma^2 = \sigma_v^2 + \sigma_u^2,$$

$$\sigma_* = \sqrt{\sigma_v^2 \sigma_u^2 / \sigma^2}, \tag{7}$$

$$u_{i*} = -\sigma_u^2 \varepsilon_i / \sigma^2,$$

and $\phi(\cdot)$ and $\Phi(\cdot)$ are the pdf and cdf of the specified distribution, respectively. To derive a more intuitive percentage indicator, one may compute the efficiency level ($\exp(-u_i)$) as per Equation (1), and then the inefficiency score $\widetilde{u}_i = [1 - \exp(-\widehat{u}_i)] * 100\%$.

The ALS77 model has been applied in multiple economic sectors after it was introduced to the literature. It also has been developed into more advanced models for a wider range of empirical purposes. The analysis by Wagstaff (1989) appears to be the first application for the efficiency of hospitals using ALS77, followed by Zuckerman et al. (1994); Rosko (1999); Chirikos and Sear (2000); Farsi and Filippini (2008) for example, as well as studies in other healthcare sectors, e.g., Vitaliano and Toren (1994); Farsi et al. (2005) for the efficiency of nursing homes.

2.2 Application of ALS77

In the context of R, one can construct some SFMs with packages for general use, e.g., the **lm** function by R Core Team (2022) and the **plm** package for panel data estimators by Croissant and Millo (2008). It is also convenient to conduct SFMs with the powerful user-written packages specifically for SFA, i.e., the **frontier** package by Coelli and Henningsen (2020), which is flexible for several SFMs, the **sfaR** package by Dakpo et al. (2021), allows for multiple types of distributions of error term, as well as the function **sfa** in the package **Benchmarking** by Bogetoft and Otto (2020), etc.

One feasible approach to apply the ALS77 model in R is by using the **frontier** package as illustrated in the snippet of code in Box 1. The model can be estimated using the function **sfa** with

```
1  ######################### ALS77 #########################
2  # Cross-sectional data frame "data" and formula of production "form" have
     been pre-defined
3  library(frontier)
4  attach(data)
5  # Model estimation
6  als77 <- sfa (form, data = data, ineffDecrease = T, truncNorm = F,
     timeEffect = F)
7  summary(als77)
8  # Individual inefficiency by JLMS
9    # Coefficients
10   als77coef <- coef(als77, which = "mle", extraPar = T)
11   # epsilon from fitted values
12   fals77 <- fitted(als77, asInData = T)
13   ei = lAggout - fals77
14   # E(ui|ei)
15   us2 = (als77coef[["sigmaU"]])^2
16   vs2 = (als77coef[["sigmaV"]])^2
17   sigmastar = sqrt((vs2*us2)/(vs2+us2))
18   ustari = (-us2*ei)/(vs2+us2)
19   uals77 = ((sigmastar*dnorm(ustari/sigmastar))/(pnorm(ustari/sigmastar)))
        +ustari
20   inals77 = 1 - exp(-uals77)
21   summary(inals77)
```

Box 1: R code snippet for ALS77

the defined production relationship, data source and distributions, etc. The individual technical inefficiency can then be estimated by the function **efficiencies**. Alternatively, in this example for ALS77, we follow the JLMS method as in Equation (6) and (7) and obtain the estimators incorporating the function **coef** and **fitted** to extract the coefficients estimation and fitted value of the production function.

For ALS77 and more SFMs, including the representative models that we will discuss later, one can also use the functional commands in Stata for the estimations. For more details, one can refer to Nguyen et al. (2022), where a wide spectrum of SFMs were introduced and instructive implementations for them were provided with the Stata code and illustrative examples. As summarized in their chapter, one can use the command sfcross by Belotti et al. (2013) or the step-wise commands in Kumbhakar et al. (2015) for flexible model specifications. Meanwhile, as introduced therein, we can also use Matlab to solve the MLE for SFMs, e.g., following the code provided in

Sickles and Zelenyuk (2019).[2]

3 SFMs with Panel Data

3.1 Schmidt and Sickles (1984)

An important direction of the extension based on the basic SFM is to formulate the model in panel data settings. One of the seminal models in the literature that extended the framework of SFM from cross-sectional data to panel data was introduced by Schmidt and Sickles (1984) (hereafter SS84) as

$$y_{it} = a + x_{it}'\beta + v_{it} - u_i, \quad i = 1,\ldots,n; \ t = 1,\ldots,T, \tag{8}$$

where $y_{it} \in \Re_+$ represents the output of unit i in time t. To save space, we omit the notation of different input variables and use vector $x_{it} \in \Re_+^p$ for the p inputs noted by unit and time, and we use vector β for their corresponding parameters. The one-side error term u_i ($u_i \geq 0, i = 1,\ldots,n$) represents the inefficiency of unit i, and is assumed to be independent from the random error v_{it}, while the latter error term v_{it} is assumed uncorrelated with the input vector x_{it}.

With simple transformation, we can rewrite Equation (8) into a common formation of the panel data model as

$$\begin{aligned} y_{it} &= a^* + x_{it}'\beta + v_{it} - u_i^*, \\ &= a_i + x_{it}'\beta + v_{it}, \end{aligned} \tag{9}$$

where

$$\begin{aligned} a^* &= a - E(u_i), \\ u_i^* &= u_i - E(u_i), \\ a_i &= a - u_i, \\ &= a^* - u_i^*. \end{aligned} \tag{10}$$

Consequently, Equation (9) can be estimated with the general methods for panel data models, e.g.,

---

[2]For interested readers, the code can be downloaded at: `https://sites.google.com/site/productivityefficiency/home`.

estimators by ordinary least squares and generalized least-squares, as well as the within estimators, the Hausman-Taylor estimators and MLE (Schmidt and Sickles, 1984). Therefore, one advantage of SS84 over the classical ALS77 and most other SFA models is the relaxed assumption on the distributions of the error terms (i.e., v_{it} and u_i), which allows more estimators than MLE alone during the model estimation. On the other hand, one constraint of SS84 is that the inefficiency represented by u_i is then assumed time-invariant, which is a limit especially in modeling panel data.

With the estimated parameters for the standard panel data model, we can obtain an estimator of the technical inefficiency ($\hat{u}_i$) as proposed in Schmidt and Sickles (1984) as

$$\widehat{u}_i = \max_i(\widehat{a}_i) - \widehat{a}_i, \quad i = 1, \ldots, n. \tag{11}$$

As a seminal model for SFA in panel data framework, the SS84 model also has been applied in numerous studies, e.g., it was also discussed in the efficiency analysis of Spanish hospitals by Wagstaff (1989), as well as in Li and Rosenman (2001) for the hospitals in Washington, the US and in Farsi et al. (2005) for the nursing homes in Switzerland. Moreover, many models have been developed based on SS84 and more applications using them have been conducted in various fields of study.

3.2   Application of SS84

An application of the SS84 model is illustrated in the snippet of R code in Box 2. We first estimate the panel data model in Equation (9) with the **plm** function, using the within estimator for fixed effect in this example. Then the fixed effect is extracted using function **fixef** to calculate the estimated inefficiency term in Equation (11).

```
1  ######################## SS84 ########################
2  # Panel data frame "paneldata" and formula of production "form" have been
       pre-defined
3  library(plm)
4  attach(paneldata)
5  # Fixed effects for ui, specifying unit id and time
6  ss84 <- plm(form, data = paneldata,
7              model = "within", index = c("HOSID","Yeardummy"), effect = "
                  individual")
8  summary(ss84)
9    # ai and uihat
10   ai = as.numeric(unname(fixef(ss84)))
11   uss84 = max(ai) - ai
12   inss84 = 1 - exp(-uss84)
13   summary(inss84)
```

Box 2: R code snippet for SS84

3.3 Pitt and Lee (1981)

The relaxed assumption of the error term distributions is one of the most notable features of the SS84 model. An example that is also in the form of a standard panel data model, is the Pitt and Lee (1981) model (hereafter PL81), which can be expressed the same as Equation (9), while adding assumptions on the distributions of v_{it} and u_i that

$$
\begin{aligned}
v_{it} &\sim iidN(0, \sigma_v^2), \\
u_i &\sim iidN^+(0, \sigma_u^2).
\end{aligned}
\tag{12}
$$

Therefore, the MLE is required for the model estimation, using the results of the standard panel regression as the initial values. Meanwhile, the JLMS approach is also available to estimate the individual inefficiency with the estimated parameters.

One may notice that the normal and half normal distributions of the noise and inefficiency term in Equation (12) are similar to the assumptions in Equation (5) for the ALS77 model. Yet there is some promotion in regard to the panel data framework on the statistical noise term (i.e., v_{it} instead of v_i), though the inefficiency term u_i is still assumed to be time-invariant.

As a result, although the application of PL81 in the healthcare sector appears to be sparse (e.g., Comans et al. (2020) for the aged care in Australia), it is still usually considered an important node

```
1  ######################### PL81 #########################
2  # Panel data frame "paneldata" and formula of production "form" have been
     pre-defined
3  library(plm)
4  library(frontier)
5  attach(paneldata)
6  # Random effects, specifying unit id and time
7  pl81f <- plm(form, data = paneldata,
8               model = "random", index = c("HOSID","Yeardummy"), effect = "
                 individual")
9  summary(pl81f)
10   # Extract initial values
11   sigmau2 = as.numeric(ercomp(pl81f)[["sigma2"]][2])
12   sigmav2 = as.numeric(ercomp(pl81f)[["sigma2"]][1])
13   sigmasq = sigmau2+sigmav2
14   gamma = sigmau2/sigmasq
15
16   init_para = c(as.numeric(pl81f$coefficients), sigmasq, gamma)
17  # MLE with initial values
18  pl81 <- sfa (form, data = paneldata, ineffDecrease = T, truncNorm = F,
     timeEffect =F,
19               startVal = init_para)
20  summary(pl81)
21
22   # Estimate individual inefficiency with user-written functions
23   inpl81 <- 1-efficiencies (pl81, asInData = T, logDepVar = T, minusU = T)
24   summary(inpl81)
```

Box 3: R code snippet for PL81

in the development routine of SFMs in panel data (e.g., Gong and Sickles (1992)).

3.4   Application of PL81

As illustrated in Box 3, the PL81 model can first be estimated with panel regression, using the **plm** function and assuming random effect. The estimated parameters are then used as the initial values in the Maximum Likelihood Estimation with function **sfa**. Consequently, the individual level of efficiency can be obtained via the **efficiencies** function with the specified estimated model.

3.5   Cornwell et al. (1990)

As indicated in Equation (9), the SS84 model and the alike PL81 model assume the inefficiency term u_i as time-invariant, which is not "perfect" in the panel data framework (though no model

can be perfect in a sense). One solution to estimate time-variant inefficiency, while sharing the advantages (e.g.., more flexible in estimator options and an easier computed inefficiency term) under the SS84 framework is the model introduced by Cornwell et al. (1990) (hereafter CSS90).

Under the framework of SS84 as in Equation (9), the CSS90 model assumes the intercept as time-variant, i.e.,

$$y_{it} = a_{it} + x'_{it}\beta + v_{it}, \quad i = 1, \ldots, n; \ t = 1, \ldots, T, \tag{13}$$

where a_{it} is a parameterized function of time t, i.e.,

$$a_{it} = \lambda_{0i} + \lambda_{1i}t + \lambda_{2i}t^2. \tag{14}$$

Consequently, the a_{it} as estimated in Equation (13) can be regressed in Equation (14). Following the method in Schmidt and Sickles (1984), the individual inefficiency at time t can then be obtained with the fitted value of Equation (14) as

$$\widehat{a}_t = \max_{j}(\widehat{a_{jt}}),$$
$$\widehat{u_{it}} = \widehat{a}_t - \widehat{a_{it}}, \quad i = 1, \ldots, n; \ t = 1, \ldots, T. \tag{15}$$

To the best of our knowledge, the CSS90 approach has not been applied in the context of hospital, yet the reader can see Comans et al. (2020) for an application of the aged care.

3.6 Application of CSS90

A snippet of R code for the CSS90 model is illustrated in Box 4. Following the same procedure of SS84, the standard panel regression with fixed effect is firstly estimated with **plm** function. The residuals are then regressed in Equation (14) by unit i. Finally, the inefficiency term can be estimated by the approach in Equation (15).

```
1  ######################### CSS90 #########################
2  # Panel data frame "paneldata" and formula of production "form" have been
      pre-defined
3  library(plm)
4  attach(paneldata)
5  # Fixed effects for ui
6  c90ori <- plm(form, data = paneldata,
7                model = "within", index = c("HOSID","Yeardummy"), effect =
                   "individual")
8  summary(c90ori)
9    # Vectors for regression with t
10   ai = as.numeric(c90ori[["residuals"]])
11   t = c(1, 2, 3, 4)
12   t2 = t^2
13   ait = numeric()
14 # Regression by unit
15 for (i in 1:length(unique(HOSID))){
16   aittemp = lm(ai[(4*(i-1)+1):(4*(i-1)+4)] ~ t2 + t)$fitted.values
17   aittemp = as.numeric(unname(aittemp))
18   ait = c(ait, aittemp)
19 }
20 c90 = cbind(ait, t)
21   # cthat for each year
22   ajtyear = aggregate(ait ~ t, data = c90, max)
23   # Compute uithat
24   c90 = as.data.frame(cbind(c90, ajtyear[,2]))
25   colnames(c90) <- c("ait","t","ajt")
26   uc90 = c90$ajt - c90$ait
27   summary(uc90)
28   inc90 = 1 - exp(-uc90)
29   summary(inc90)
```

Box 4: R code snippet for CSS90

3.7 Kumbhakar (1990) and Battese and Coelli (1992)

An extension to the PL81 model would be to consider the time-variant inefficiency (i.e., u_{it} instead of u_i) in the panel data framework, while keeping the feasible distribution assumptions on the error terms. Kumbhakar (1990) (hereafter K90) and Battese and Coelli (1992) (hereafter BC92) are two widely applied models in this context, which can be formulated similar to PL81 as

$$
\begin{aligned}
&y_{it} = a + x'_{it}\beta + v_{it} - u_{it}, \quad i = 1,\ldots,n; \ t = 1,\ldots,T, \\
&v_{it} \sim iidN(0, \sigma_v^2), \\
&u_{it} = h(t)u_i, \\
&u_i \sim iidN^+(\mu, \sigma_u^2),
\end{aligned}
\tag{16}
$$

where component $h(t)$ in K90 is assumed as

$$
h(t) = [1 + \exp(bt + ct^2)]^{-1},
\tag{17}
$$

and in BC92 as

$$
h(t) = \exp[-d(t-T)],
\tag{18}
$$

where T is the end time of the period.[3] As a result, $h(t)$ is identical for every unit i in both cases.

Consequently, with the constraints of parametric assumptions, K90 and BC92 need to be estimated via MLE, and the inefficiency could then be estimated with the JLMS method.

The K90 and BC92 models are meaningful attempts in modeling time-variant inefficiency in SFA. In the context of hospital efficiency analysis, they were considered, for example, in Linna (1998); Besstremyannaya (2011).

---

[3] Another akin model is Kumbhakar and Wang (2005), where the component influencing the inefficiency is assumed as $h(t) = exp[-g(t-i)]$, where i is the beginning time of the period.

```
1 ####################### BC92 #######################
2 # Panel data frame "paneldata" and formula of production "form" have been
    pre-defined
3 library(frontier)
4 attach(paneldata)
5 # Model estimation
6 bc92 <- sfa (form, data = paneldata, ineffDecrease = T, truncNorm = T,
    timeEffect = T)
7 summary(bc92, effic = F, logDepVar = T, effMinusU = T)
8
9   # Estimate individual inefficiency with user-written functions
10   inbc92 <- 1-efficiencies (bc92, asInData = T, logDepVar = T, minusU = T)
11   summary(inbc92)
```

Box 5: R code snippet for BC92

3.8 Application of BC92

Taking BC92 as an example, a snippet of R code is illustrated in Box 5. The model can be estimated using function **sfa** with predefined distributions and time effect. Accordingly, the individual inefficiency can be then estimated with function **efficiencies**.

3.9 'True fixed effect' models

In the effort of introducing the time-variant inefficiency term in SFMs, whether the improvement of CSS90 over SS84 or that of BC92 over PL81, the individual inefficiency is not indeed separated from the time-invariant individual heterogeneity. One advanced model for such purpose is the so-called 'true fixed effect' model proposed by Greene (2005) (hereafter G05) as

$$
\begin{aligned}
y_{it} &= a_i + x'_{it}\beta + v_{it} - u_{it}, \quad i = 1,\ldots,n; \ t = 1,\ldots,T, \\
v_{it} &\sim iidN(0,\sigma_v^2), \\
u_{it} &\sim iidN^+(0,\sigma_u^2).
\end{aligned}
\tag{19}
$$

To estimate the G05 model, Greene (2005) also proposed the maximum likelihood dummy variable estimators. Eventually, the inefficiency estimation of each unit i can vary across time, while is also insensitive to the outliers in the sample. However, G05 is usually computationally

14

```
1  ######################## G05 (True fixed)########################
2  # Panel data frame "paneldata" and formula of production (with "true fixed
       effect") "formt" have been pre-defined
3  library(frontier)
4  attach(paneldata)
5  # Model estimation
6  g05 <- sfa (formt, data = paneldata, ineffDecrease = T, truncNorm = F,
       timeEffect = T)
7  summary(g05, effic = F, logDepVar = T, effMinusU = T)
8
9    # Estimate individual inefficiency with user-written functions
10   ing05 <- 1-efficiencies (g05, asInData = T, logDepVar = T, minusU = T)
11   summary(ing05)
```

Box 6: R code snippet for G05

intensive, especially when the number of units (i.e., n) is large, and is often accompanied with the incidental parameters problem (Greene, 2005; Nguyen et al., 2022). Nevertheless, the G05 model, as a recent development to distinguish transitory inefficiency, has been applied in the studies of healthcare efficiency, e.g., Herr (2008) for the hospitals in Germany, and Comans et al. (2020) for the cost efficiency of Australian aged care.

3.10 Application of G05

An illustration of the G05 model is provided in Box 6. In addition to the panel data and distribution settings in function **sfa**, the factor for dummy variables needs to be specified in the formula in advance.[4] The function **efficiencies** is also capable of estimating the individual inefficiency of the G05 model afterwards.

3.11 Kumbhakar et al. (2014)

Based on G05, another more recently developed model that includes both transitory and persistent inefficiency components was introduced by Kumbhakar et al. (2014) (hereafter KLH14),[5] which

---

[4]For more details including the formula settings in this chapter, see Appendix A.

[5]In comparison, G05 considers transitory inefficiency only.

can be modeled as

$$y_{it} = a + x'_{it}\beta + v_{it} - u_{it} + s_i - \eta_i, \quad i = 1,\ldots,n; \ t = 1,\ldots,T,$$

$$v_{it} \sim iidN(0,\sigma_v^2),$$

$$u_{it} \sim iidN^+(0,\sigma_u^2), \tag{20}$$

$$s_i \sim iidN(0,\sigma_s^2),$$

$$\eta_i \sim iidN^+(0,\sigma_\eta^2),$$

where u_{it} is transitory inefficiency now and η_i represents the persistent inefficiency of unit i, s_i is the separated individual heterogeneity, while v_{it} still represents the random error. Similar to the transformation for SS84, KLH14 can be rewritten into

$$y_{it} = a^* + x'_{it}\beta + a_i + \varepsilon_{it}, \quad i = 1,\ldots,n; \ t = 1,\ldots,T, \tag{21}$$

in which

$$a^* = a - E(\eta_i) - E(u_{it}),$$

$$a_i = s_i - \eta_i + E(\eta_i), \tag{22}$$

$$\varepsilon_{it} = v_{it} - u_{it} + E(u_{it}).$$

Consequently, one can first estimate a_i and ε_{it} with the standard panel data model in Equation (21). Then we can estimate the persistent inefficiency of unit i (i.e., $E(\eta_i)$) and the transitory inefficiency of unit i at time t (i.e., $E(u_{it})$) with the estimated $\widehat{a_i}$ and $\widehat{\varepsilon_{it}}$ applying in a standard stochastic frontier analysis, respectively. Colombi et al. (2014) (hereafter CKMV) proposed the same model as KLH14 in their study, while they suggested a single-stage MLE method for estimation. In comparison, the two-stage KLH14 model is usually easier in implementation and computation. Yet, to distinguish the persistent and transitory inefficiency, the two models also come with the cost of four distributional assumptions on error terms.

To the best of our knowledge, these recently developed approaches that separate transitory and persistent inefficiency (i.e., KLH14 and CKMV) have not been applied in the context of healthcare.

Yet, the four-component error term framework inspired many latter works for this topic, e.g., the model designed by Colombi et al. (2017) based on the framework of CKMV, which was applied in the analysis of hospital efficiency in Italy, and the maximum simulated likelihood estimator proposed by Filippini and Greene (2016), which was later used in an analysis of nursing home efficiency in Switzerland by Filippini et al. (2021).

3.12 Application of KLH14

The two-step procedure of KLH14 can be illustrated as in Box 7. The standard panel regression is first estimated with function **plm**, and a_i and ε_{it} are then estimated from the production model. In the second stage, standard stochastic frontier technique is applied to the estimated $\widehat{a_i}$ and $\widehat{\varepsilon_{it}}$ with function **sfa**, respectively.

4   SFMs with Determinant of Inefficiency

Aside from the extensions of incorporating panel data analysis in SFMs, another notable question is how the exogenous variables (i.e., those usually referred to as 'environmental factors') influence the (in)efficiency of the production units. More importantly, identifying the determinant factors and their influencing patterns to the (in)efficiency term would be meaningful for realistic practice.

4.1   Kumbhakar et al. (1991), Battese and Coelli (1995) and Caudill et al. (1995)

The model introduced in Kumbhakar et al. (1991) (hereafter KGM91) is a classical solution in the literature that includes determinants by defining the inefficiency term u_{it} with parameterized

```
1  ######################### KLH14 #########################
2  # Panel data frame "paneldata" and formula of production "form" have been
       pre-defined
3  library(plm)
4  library(frontier)
5  attach(paneldata)
6  # Fixed effects for alphai and epsilonit
7  klh14ori <- plm(form, data = paneldata,
8                  model = "within", index = c("HOSID","Yeardummy"), effect = "
                       individual")
9  summary(klh14ori)
10 # alphai
11 alphai = as.numeric(unname(fixef(klh14ori)))
12 # epsilonit
13 epsiloni = as.numeric(klh14ori[["residuals"]])
14 summary(alphai)
15 summary(epsiloni)
16 # Constant for sfa estimation
17 constant = 1
18 # Persistent inefficiency
19 klh14per = as.data.frame(cbind(alphai,constant))
20 klh14p <- sfa(alphai~constant-1, data = klh14per, ineffDecrease = T,
       truncNorm = F, timeEffect = F)
21 summary(klh14p, effic = F, logDepVar = T, effMinusU = T)
22
23   # Estimate individual inefficiency with user-written functions
24   inklh14p <- 1-efficiencies (klh14p, asInData = T, logDepVar = T, minusU
         = T)
25   summary(inklh14p)
26
27 # Transitory inefficiency
28 klh14tran = as.data.frame(cbind(epsiloni,constant))
29 klh14t <- sfa(epsiloni~constant-1, data = klh14tran, ineffDecrease = T,
       truncNorm = F, timeEffect = T)
30 summary(klh14t, effic = F, logDepVar = T, effMinusU = T)
31
32   # Estimate individual inefficiency with user-written functions
33   inklh14t <- 1-efficiencies (klh14t, asInData = T, logDepVar = T, minusU
         = T)
34   summary(inklh14t)
```

Box 7: R code snippet for KLH14

determinant variables, which can be modeled as

$$y_{it} = a + x'_{it}\beta + v_{it} - u_{it}, \quad i = 1, \dots, n,$$

$$v_{it} \sim iidN(0, \sigma_v^2),$$

$$(23)$$

$$u_{it} \sim iidN^+(\tau_{it}, \sigma_u^2),$$

$$\tau_{it} = z'_{it}\gamma,$$

where the mean of inefficiency u_{it}, compared to the half-normal distribution in ALS77, is assumed as τ_{it}, which is 'determined' by $z_{it} \in \mathfrak{R}^m$, a vector of m environmental variables, and γ, the vector of corresponding parameters.

The specification of the KGM91 model may be more well-known as the Battese and Coelli (1995) model (hereafter BC95) in the panel data framework. Meanwhile, another similar specification is the Caudill et al. (1995) model (hereafter CFG95), where the distribution of u_i is assumed as

$$u_i \sim iidN^+(0, \sigma_{u_i}^2),$$

$$(24)$$

$$\ln(\sigma_{u_i}^2) = z'_i\gamma,$$

that the environmental factors z_i influence the inefficiency via the assumed variance $\sigma_{u_i}^2$.[6]

The KGM91/BC95 models provide a convenient framework of modeling the determinant factors of inefficiency in SFA, which is an attractive feature for efficiency analysis. As a result, numerous studies have applied them in various fields of research, e.g., Rosko (2001, 2004); Herr (2008); Vitikainen et al. (2010); Varabyova and Schreyögg (2013); Wang and Zelenyuk (2022) for the hospital performance, and Comans et al. (2020) for the cost efficiency of aged care.

4.2 Application of KGM91/BC95

In this illustration, we provide an example of the KGM91/BC95 models in Box 8. In the parameter estimation by function **sfa**, one needs to specify the environmental factors in the formula in

---

[6]For further extension in this vein, see e.g., Wang and Schmidt (2002), where the truncated mean in BC95 and variance in CFG95 are both modeled in the inefficiency distribution.

```
1  ######################## BC95 #########################
2  # Panel data frame "paneldata" and formula of production (with
       environmental variables) "formz" have been pre-defined
3  library(frontier)
4  attach(paneldata)
5  # Model estimation
6  bc95 <- sfa(formz, data = paneldata, ineffDecrease = T, truncNorm = T,
       timeEffect = F)
7  summary(bc95, effic = F, logDepVar = T, effMinusU = T)
8
9    # Estimate individual inefficiency with user-written functions
10   inbc95 <- 1-efficiencies (bc95, asInData = T, logDepVar = T, minusU = T)
11   summary(inbc95)
```

Box 8: R code snippet for KGM91/BC95

advance.[7] Then, the individual inefficiency can be estimated with the **efficiencies** function.

5 Semi-parametric SFMs

The parametric assumptions in SFMs of both the production function and error terms are widely concerned as a source of limitation in the literature. So far, we have introduced models trying to relax the assumptions on the error term (i.e., the SS84 and CSS90 models), while effort has also been devoted to exploring the relaxation of the parametric assumption on the production function, or on both of them simultaneously.

One stream of the so-called 'semi-parametric' approach was first introduced by Fan et al. (1996), where the non-parametric kernel regression is incorporated with the parametric MLE in estimating the production relationship and frontier. In a more recently developed non-parametric framework introduced by Kumbhakar et al. (2007), the local maximum likelihood estimator (LMLE) was suggested to replace MLE, where the weight of likelihood is based on a kernel function of the local variables. Furthermore, Simar et al. (2017) (hereafter SVKZ) proposed using non-parametric least squares instead of LMLE to further reduce the assumption requirements. We will focus on the SVKZ algorithm to illustrate the semi-parametric paradigm in the following subsection.

---

[7]For more details including the formula settings, see Appendix A.

5.1 Simar et al. (2017)

The SVKZ model formulates output $y_i \in \Re_+$, inputs vector $x_i \in \Re_+^p$ and vector of m influential variables, $z_i \in \Re^m$ as

$$y_i = g(x_i, z_i) + v_i - u_i, \quad i = 1, \ldots, n, \tag{25}$$

where $g(x_i, z_i)$ represents the production frontier, the random error v_i is assumed as a real random distribution that $v_i \sim D(0, \sigma_{v_i}^2(x_i, z_i))$, and the inefficiency term u_i is assumed as a positive random distribution that $u_i \sim D^+(\mu_{u_i}(x_i, z_i), \sigma_{u_i}^2(x_i, z_i))$. Meanwhile, v_i and u_i are assumed as independent random variables conditionally on (x_i, z_i). Specially, when defining $\varepsilon_i = v_i - u_i + \mu_{u_i}(x_i, z_i)$ and $r_1(x_i, z_i) = g(x_i, z_i) - \mu_{u_i}(x_i, z_i)$, Equation (25) can be rewritten as

$$y_i = r_1(x_i, z_i) + \varepsilon_i, \tag{26}$$

while based on distribution assumptions of v_i and u_i above, it can be derived that

$$E(\varepsilon_i | x_i, z_i) = 0, \tag{27}$$

$$E(\varepsilon_i^2 | x_i, z_i) = \sigma_{v_i}^2(x_i, z_i) + \sigma_{u_i}^2(x_i, z_i), \tag{28}$$

$$E(\varepsilon_i^3 | x_i, z_i) = -E[(u_i - \mu_{u_i}(x_i, z_i))^3 | x_i, z_i]. \tag{29}$$

Consequently, if we define $r_j = E(\varepsilon_i^j | x_i, z_i)$, according to Equation (26), we can then estimate r_1 through nonparametric methods, such as local least squares, and then estimate r_2 and r_3 with nonparametric methods using r_1, based on the relationship in Equation (26) that $\varepsilon_i = y_i - r_1(x_i, z_i)$.

To estimate the inefficiency of individual unit i, assumption on the distribution of inefficiency u_i is required as $u_i \sim N^+(0, \sigma_{u_i}^2(x_i, z_i))$, while the assumption on the distribution of error term v_i is

not needed. Accordingly, we can further derive Equation (28) and (29) as

$$r_2(x_i, z_i) = \sigma_{v_i}^2(x_i, z_i) + (\frac{\pi - 2}{\pi})\sigma_{u_i}^2(x_i, z_i), \tag{30}$$

$$r_3(x_i, z_i) = \sqrt{\frac{2}{\pi}}(\frac{\pi - 4}{\pi})\sigma_{u_i}^3(x_i, z_i), \tag{31}$$

so that with the estimated $\widehat{r_3}(x_i, z_i)$, we can obtain an estimation of σ_{u_i} as

$$\widehat{\sigma_{u_i}}^3(x_i, z_i) = \sqrt{\frac{\pi}{2}}(\frac{\pi}{\pi - 4})\widehat{r_3}(x_i, z_i), \tag{32}$$

and further estimate σ_{v_i} with the estimated $\widehat{r_2}(x_i, z_i)$ and $\widehat{\sigma_{u_i}}(x_i, z_i)$. Finally, the conditional mean of inefficiency $\mu_{u_i}(x_i, z_i)$ can be derived by

$$\widehat{\mu_{u_i}}(x_i, z_i) = \sqrt{\frac{2}{\pi}}\widehat{\sigma_{u_i}}(x_i, z_i). \tag{33}$$

In addition, with the definition of r_1 above, the stochastic frontier can be estimated as

$$\widehat{g}(x_i, z_i) = \widehat{r_1}(x_i, z_i) + \widehat{\mu_{u_i}}(x_i, z_i). \tag{34}$$

For further developments, interested readers can find more discussions of various routines for statistical tests of SVKZ and other semi- and non-parametric models in Parmeter and Zelenyuk (2019).

One application of this recently suggested semi-parametric method is in the context of aged care by Comans et al. (2020). Nevertheless, the class of semi-parametric models has a great potential in the further development of the efficiency analysis in general and for the healthcare field in particular.

5.2 Application of SVKZ

In Box 9, we provide a snippet of code to apply the SVKZ algorithm in R, which is modified based on the programming by Parmeter and Zelenyuk (2019). The non-parametric kernel regression is computed in function **npreg** in package **np** by Hayfield and Racine (2008). Besides, the formulas need to be specifically defined for Equation (26). Meanwhile, a comprehensive set of implementation methods in Stata for SVKZ, as well as for the other models in this chapter, can be found in Nguyen et al. (2022).

6 Empirical Illustration of Selected Popular SFA Models

6.1 Sample Data

In the following section, we apply the series of selected popular SFMs briefly introduced above to a real-world data set as an empirical illustration. The data comes from Queensland Health, and contains operational records of 95 public hospitals in Queensland, Australia during the period from Financial Year (FY) 2012/13 to FY 2015/16 (i.e., 380 observations). In this preliminary analysis, following the recent studies of the Queensland hospital efficiency (e.g., Nguyen and Zelenyuk (2021a,b); Wang and Zelenyuk (2022)) as well as the practice for hospitals in other regions (e.g., Grosskopf and Valdmanis (1987); Rosko (2001) for the US), we select three inputs (i.e., labor, expenditure, beds) and one output (i.e., the aggregated inpatient and outpatient services) for the production function.[8] Besides, we consider three environmental variables for the models considering determinants of inefficiency, i.e., the location, teaching status and size. The descriptive statistics of the selected variables are summarized in Table 1.

---

[8]Similar to the data processing in Nguyen and Zelenyuk (2021a), the input variables of labor and output variables are aggregated respectively with a Principal Component Analysis (PCA) based approach proposed by Daraio and Simar (2007).

```r
####################### SVKZ #########################
library(np)
attach(data)
# Define factor variables and formulas
fTEACH = as.factor(TEACH)
fRemote = as.factor(Remote)
fSmall = as.factor(Small)

forms = lAggout ~ lBEDS + lAgglabours + lSUPP + fTEACH + fRemote + fSmall
forms3 = ehat3 ~ lBEDS + lAgglabours + lSUPP + fTEACH + fRemote + fSmall

# Bandwidths selection
bws.r1 <- npregbw(forms, regtype="ll", data=data, bwmethod = "cv.ls",
    ckertype = "epanechnikov")
# Estimate conditional mean and extract fitted values and residuals
r1.est <- npreg(bws=bws.r1, gradients=TRUE)
r1hat   <- fitted(r1.est)
e1hat   <- residuals(r1.est)
# Generate for r3
ehat3 <- e1hat^3

# Bandwidths selection for r3 (skewness measures)
bws.r3 <- npregbw(forms3, data=data, regtype="ll", bwmethod = "cv.ls",
    ckertype = "epanechnikov")
# Estimate conditional mean and extract fitted values and residuals
r3.est  <- npreg(bws=bws.r3, gradients=TRUE)
r3hat   <- fitted(r3.est)
e3hat   <- residuals(r3.est)

  # Estimate individual inefficiency
  sigu3.hat <- sqrt(pi/2)*(pi/(pi-4))*r3hat
  sigu.hat <- apply(cbind(0,sigu3.hat),1,FUN=max)^(1/3)
  muhat <- sqrt(2/pi)*sigu.hat
  summary(muhat)
  insvkz = 1 - exp(-muhat)
  summary(insvkz)
```

Box 9: R code snippet for SVKZ

Table 1: Descriptive statistics of Queensland public hospitals, FY 2012/13 to FY 2015/16

Variable	Description	Mean	Std Dev	Min	Max
Input					
Agglabours	Aggregated labor input	0.76	1.57	0.01	8.71
SUPP*	Consumable expenditure	7.83	19.20	0.03	164.00
BEDS	Number of beds	74.92	133.78	3.00	680.00
Output					
Aggout	Inpatient and outpatient service	0.54	1.03	0.01	5.05

Variable	Description	Frequency	Percentage
Environmental			
Remote	Located in remote areas	108	28.42%
Non-remote	Located in non-remote areas	272	71.58%
Small	Small hospitals	300	78.95%
Big	Large hospitals	80	21.05%
TEACH	Teaching hospitals	70	18.42%
Non-teaching	No teaching function	310	81.58%

* AU$1,000,000 in constant price of FY2012/2013

6.2 Estimations of Inefficiency

All the models in this brief illustration are estimated in a linear and logarithmic Cobb-Douglas production function. Alternative production functions, such as the Translog function, the Leontief function, the constant elasticity substitution (CES) function, etc. are also optional in the SFA context. According to the results of the model estimation, the individual level of inefficiency is then estimated within each model. As a result, the descriptive statistics of the inefficiency level are summarized in Table 2.

In addition, to visualize the distribution, we plot the estimated densities of the inefficiency level estimated by each model in Figure 1, using a nonparametric kernel density estimator with Gaussian kernel, where the bandwidth is selected by cross-validation.[9]

Intuitively, the estimated inefficiency scores of different models vary dramatically. This phenomenon is as expected since the assumptions on the distributions of error terms and even the production function of different models are quite different. The variation is still significant even

---

[9]Full R code for the analysis and visualization is in Appendix A.

Table 2: Statistics of the estimated level of inefficiency (%) by different models

Model	Mean	Std. Dev.	95% CI of mean	Min	Q1	Median	Q3	Max
ALS77	19.97%	6.39%	[19.33%, 20.62%]	6.94%	15.63%	18.97%	22.59%	47.37%
SS84	82.22%	24.61%	[77.21%, 87.24%]	0.00%	84.55%	92.64%	96.61%	98.71%
PL81	46.45%	22.22%	[44.21%, 48.69%]	2.44%	26.35%	44.28%	65.13%	83.75%
CSS90	25.97%	15.94%	[24.36%, 27.58%]	0.00%	15.22%	18.99%	37.42%	62.44%
BC92	58.34%	15.78%	[56.75%, 59.93%]	4.66%	48.03%	57.37%	69.65%	87.37%
G05	0.86%	1.53%	[0.71%, 1.02%]	0.00%	0.00%	0.00%	0.62%	6.08%
KLH14T	0.09%	0.00%	[0.09%, 0.09%]	0.08%	0.09%	0.09%	0.09%	0.09%
KLH14P	0.39%	0.00%	[0.39%, 0.39%]	0.39%	0.39%	0.39%	0.39%	0.39%
BC95	30.76%	21.10%	[28.63%, 32.89%]	3.58%	11.45%	27.28%	47.58%	78.11%
SVKZ	8.78%	10.55%	[7.71%,9.84%]	0.00%	0.00%	0.00%	19.66%	33.03%

between some models of similar specifications, e.g., most of the inefficiency levels estimated in the SS84 model are very high with a mean of 82.22%, while the estimation by the CSS90 model is relatively lower with an average of 25.97%. Therefore, as indicated in the analysis with different SFMs in different production functions by Wang and Zelenyuk (2022), selecting the suitable models and corresponding assumptions for the research question is an important foundation of analysis, especially in the case of proposing policy suggestions.

In comparison, the results of the PL81 and BC92 models are relatively close to each other, ranging from around 2% to 84% and 5% to 87%, respectively. The mean inefficiency level estimated by these two models are around 50%, while the standard deviations are relatively high. One may notice that the standard deviations of several models are also relatively high on the basis of their means, such as those of CSS90, BC95 and SVKZ. Nevertheless, the confidence intervals of the mean inefficiency of most models are narrow, benefiting from the relatively large data set.

We also observe that the models separating the inefficiency from the individual heterogeneity (i.e., the G05 and KLH14 models) estimate a quite high level of efficiency for most observations. Such a degeneracy phenomenon, again, emphasizes the importance of selecting the appropriate model in SFA, incorporating the fitted production function and other model specifications.

The inefficiency level estimated in the basic ALS77 model is lower than the majority of other models, ranging from 7% to 47% with an average of about 20%. Besides, the standard deviation

of ALS77 is also among the lowest of about 6%, so that one can also observe a smoother tendency of the kernel density estimation for ALS77.

So far in the literature, several studies have been conducted for the hospital efficiency in Queensland, especially with the SFA techniques. For example, Nghiem et al. (2011) applied SFA with Malmquist Productivity Index (MPI) to evaluate the total factor productivity (TFP) of public hospitals during 1996 to 2004. O'Donnell and Nguyen (2013) conducted a conventional SFA model with a wider group of public hospitals during the same period. Moreover, the analysis of Australian public hospitals between FY 1995/96 and FY 2005/06 by Gabbitas and Jeffs (2009) may be akin to our illustration here, where PL81 and BC95 were applied for random effects and SS84 and G05 were applied for fixed effect. In comparison, their estimated mean inefficiency levels for Queensland in particular, are relatively lower than those in our sample, which were reported as about 20% with the PL81, BC95 and SS84 models, and 8% with the G05 model.
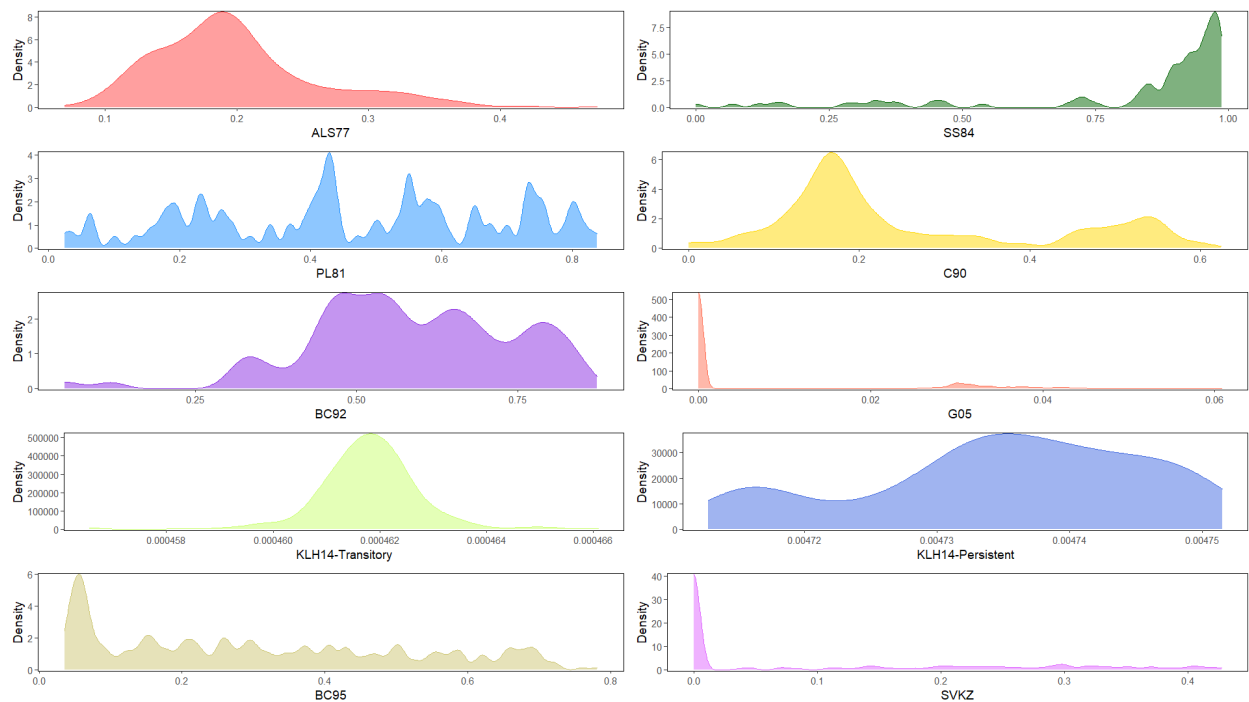

Figure 1: Estimated kernel densities of the inefficiency level of different SFMs

7 Other SFA Approaches

The goal of this section is to briefly review some of the other SFA approaches that have been rarely applied to the context of healthcare, yet we believe have good potential for future research endeavors in this area. These include some of the most recent advances in SFA. Our main focus will be on spacial SFA models, which we will cover in the first sub-section, while in the final sub-section we will briefly mention some classic and recent works on Bayesian SFA.

7.1 The Use of Spatial Stochastic Frontier Methods in Modeling Health Outcomes

In their very recent study "Homelessness on the West Coast: What Role Does Health Play?" Fuller and Sickles (2022) utilize spatial stochastic frontier methods to analyze the impacts of health shocks (e.g., the recent COVID pandemic) on homelessness and health. Such methods were introduced into the efficiency and productivity literature in a series of papers by Glass, et al.[10] and are also discussed in Sickles and Zelenyuk (2019). In this section, we use the recent study by Fuller and Sickles (2022) to highlight various approaches that one can take to introduce health outcomes into models in which *spillovers* play an important role, especially when health shocks such as pandemics display such obvious contagion impacts that have strong spatial dimensions.

Homelessness is a complicated social problem, due to the vast array of socioeconomic factors that influence housing status. In America, homelessness has been rising since 2017, with this increase being driven primarily by the growth of unsheltered homeless populations in West Coast cities such as Los Angeles, San Francisco, and Seattle.

Fuller and Sickles (2022) examine the socioeconomic variables influencing homelessness on the West Coast in recent years, utilizing panel fixed effects models that explicitly include measures of healthcare access and availability to account for the additional health risks faced by individuals who lack shelter. They estimate a spatial error model in order to examine potential geographic

---

[10]I.e., Glass et al. (2013), Glass et al. (2014), Glass et al. (2016a) and Glass et al. (2016b).

dependencies. In light of the dual health and economic threats posed by the recent COVID pandemic, their results are then used to test the impact and diffusion of a systemic shock, such as the pandemic and find evidence of linkages among income inequality, housing supply, healthcare investment, and homelessness as well as significant and negative spatial relationships.

Their theoretical model predicts that in regions where healthcare access is worse more people would enter into homelessness over time due to the heightened negative consequences of being homeless, which leads to a decline in future earnings potential. In order to analyze the relationship between homelessness and measures of healthcare access, income, and demographic traits, they first examine variations on a general least squares model where the dependent variable (their measure of homelessness) changes. The second part of their analysis extends this model to include a spatial component, accounting for potential dependencies between nearby Continuum of Cares (CoCs) and how that may impact homelessness levels across the West Coast.

They utilize a fixed effect model based on the framework of SS84 and CSS90 (discussed above) wherein the composed error term contains the idiosyncratic error (v_{it}) and a term that reflects time-varying heterogeneity (a_{it}) that can be interpreted as a lower bound on homelessness. Data is largely drawn from the United States Health Resources and Services Administration's (HRSA) Area Health Resource File, which is reported for the years 2010-2017.

The general fixed effects model that they estimate is:

$$HL_{it} = a_{it} + \beta_1 Grnt_{it} + \beta_2 Gini_{i,t-2} + \beta_3 MR_{it} + \beta_4 H_{i,t-2} + \beta_5 D_{i,t-1} + \beta_6 X_{it} + v_{it}, \qquad (35)$$

where the dependent variable HL_{it} is the number of homeless individuals per 1000 in a CoC in a given year.[11] $Grnt_{i,t}$ represents HRSA grant funding per person (in 2010 dollars), $Gini_{i,t-2}$ represents the Census Bureau's Gini Index calculation, and MR_{it} is the median rent value for a two-bedroom housing unit reported by the Department of Housing and Urban Development (HUD) in the US. $H_{i,t-2}$ is a matrix of health and healthcare variables, including physicians per 1000 people, the percentage of uninsured individuals, and opioid prescriptions per 100 people. $D_{i,t-1}$ is a matrix

---

[11]Three measures are used in Fuller and Sickles (2022), but for this illustration, we will use their first measure.

of demographic variables, including the number of housing units per 100 people, as well as the proportions of the population in a CoC that are black, male, and older than 65. Lastly, X_{it} is a matrix of interaction terms between the various socioeconomic variables already present in the regression. These interaction terms are simple products between two variables.

Furthermore, Fuller and Sickles (2022) examine the spatial dependencies between adjacent CoCs using both spatial autoregressive (SAR) and spatial error (SEM) models. The SAR is formulated as

$$HL_{it} = a_{it} + \beta_1 Grnt_{it} + \beta_2 Gini_{i,t-2} + \beta_3 MR_{it} + \beta_4 H_{i,t-2} + \beta_5 D_{i,t-1} + \beta_6 X_{it} + \rho \sum_{j=1}^{n} w_{ij} HL_{it} + v_{it},$$

$$(36)$$

while the SEM is

$$HL_{it} = a_{it} + \beta_1 Grnt_{it} + \beta_2 Gini_{i,t-2} + \beta_3 MR_{it} + \beta_4 H_{i,t-2} + \beta_5 D_{i,t-1} + \beta_6 X_{it*} + v_{it}, \qquad (37)$$

where

$$v_{it} = \lambda_m \sum_{j=1}^{n} w_{ij} v_{jt} + \pi_{it}. \qquad (38)$$

In this section we focus on the spatial aspects of the health/homelessness nexus. More recent studies of spatial relationships that focus on general efficiency/productivity impacts in spillover models are also available.[12]

The spatial component of the SAR model is a spatially lagged dependent variable. In this context, it evaluates whether homelessness in each CoC is influenced by homelessness in nearby CoCs. The SEM model includes the spatial term in the error component. This is meant to correct for potential spatial autocorrelation, notwithstanding any underlying spatial relationship between CoCs. For both models, the way that the spatial weights matrix W is defined will impact the results. They estimate two different versions: one with the weight matrix formed using a k-nearest-neighbors approach $(k = 5)$, and the other with the weights matrix formed using the inverse distance-squared

---

[12]See, for example, Liu and Sickles (2021); Liu et al. (2022b,a)

approach. Weighting matrices are row-standardized, and created using **GeoDa** software.

Both the spatial autoregressive model (SAR) and spatial error model (SEM) produce similar results. In the SAR model, the spatial component, ρ, is negative and statistically significant in the cases where total homelessness and unsheltered homelessness per 1000 are the variables of interest. Rather than implying that homelessness in one region exerts a positive influence on homelessness in adjacent regions, this negative coefficient suggests that homelessness levels in nearby regions are dissimilar to one another. This could be due to homeless populations migrating or clustering in specific cities or regions, but it suggests that rising (falling) homeless populations in one CoC would not lead to a similar increase (decrease) in neighboring ones.

The finding of negative spatial dependence also suggests that shocks to variables outside of our model, or outside of the level of homelessness itself, tend to affect homelessness levels in nearby regions in different ways. Further examination of this conclusion can be analyzed by examining the spatial effects of unforeseen shocks to SEM error.

In a SEM framework, spatial dependence can be influenced by factors other than the spatially lagged dependent variable (Glass et al., 2012). In the context of the homeless crisis on the West Coast, the outbreak of coronavirus across the globe is an example of an unexpected, exogenous shock. This could influence sudden changes in homelessness across regions, and it is possible to use the results from the spatial error model to examine the diffusion of such a shock across geographic regions. To do this, a multiple standard deviation change in the error component for each CoC is tested to see how that shock would affect homelessness across regions.

Recalling Equation (37) for the SEM model, one can consider impacts on the most recent counts of homelessness for 2019. The impact is measured as the change in the dependent variable resulting from a shock to the error term, π_{it}. To represent this shock, a constant term is added to the error term which takes on one of the two values: 2σ or 5σ. Here, σ is the standard deviation of the idiosyncratic error term. These two different values are used to illustrate varying degrees of a shock's severity. Thus, in the event of a shock, the error term for each CoC takes on the value of $\pi_{it} + 2\sigma$ or $\pi_{it} + 5\sigma$. The new values of the error terms then are used to determine the new value

of the dependent variable within every geographic unit. Thus, one can create an $(n \times n)$ matrix Y^*, where column j of Y^* represents the new values of the dependent variable in every CoC resulting from the shock occurring in the jth CoC.

In order to evaluate the impacts of a shock on each CoC, one can create a matrix $D = (Y^* - Y)$, where Y is a $(n \times n)$ matrix such that each column lists the values of the dependent variable for all n CoCs absent a shock. The contagion effects are constructed by forming a matrix V which is created by dividing each element of D by the corresponding element of the matrix Y. Thus, in summary, the following matrix operations are conducted:

$$Y^* = a + \beta X + (I - \lambda W)^{-1} \text{diag}(b\sigma), \tag{39}$$

$$D = Y^* - Y \tag{40}$$

$$= (I - \lambda W)^{-1} \text{diag}(b\sigma), \tag{41}$$

$$[V]_{ij} = \frac{[D]_{ij}}{[Y]_{ij}}. \tag{42}$$

Note that a is the fixed effects vector, X is the matrix of explanatory variables, W represents the spatial weights matrix from the spatial error model, and $b = 2$ or 5, depending upon the size of the shock. Thus, the values of the matrix V can be interpreted in the following way, as stated by Le Gallo et al. (2005). The diagonal elements of V represent the direct impact of a shock, occurring in a CoC, on the CoC itself. Each column of V represents the impact that a shock in one CoC has on every other CoC, and each row of V represents how each CoC is impacted by shocks in every other CoC.

Consequently, the simulated shocks in major cities predict a decrease in homelessness in nearby geographic areas, suggesting that a shock in one area could signal nearby areas to expand outreach and take preventative action in the face of a crisis.

To conclude this sub-section, it is worth noting that while we have described this type of SFA approach in the context of analyzing the impacts of health shocks on homelessness and health (following Fuller and Sickles (2022)), the possible range of other applications of this method to

healthcare is much wider and we hope this section will encourage such research.

7.2 Bayesian SFA

In 1997, Koop et al. (hereafter KOS97) first introduced Bayesian inferences into unit-specific efficiency analysis based on SFMs in panel data settings, i.e., under the classical framework in Equation (8) as

$$y_{it} = a + x_{it}'\beta + v_{it} - u_i, \quad i = 1,\ldots,n; \ t = 1,\ldots,T. \tag{43}$$

Two Bayesian frameworks for the fixed and random effect models were developed in KOS97, respectively. The unit-specific efficiency in the fixed effect models is measured relatively to the most efficient unit, which is similar to the method in the classical SS84 model as introduced above. Specifically, they proposed two fixed effect models: the *standard individual effects model* and the *marginally independent efficiency distribution model*, where the prior for $a_i = a - u_i$ in the second model is informative on the unit-specific effects. In the random effect models, an informative prior for u_i is considered so that the absolute efficiency rather than the relative efficiency is estimated. Specifically, two random effect models were proposed: the *varying efficiency distribution model* and the *common efficiency distribution model*, where the priors are determined by the features of the unit or obtained from a certain distribution, respectively.

Other than the seminal KOS97 models, more effort has been devoted to the Bayesian stochastic frontier analysis during the 1990s, including a series of literature mainly published by Koop, Osiewalski and Steel (Kim and Schmidt, 2000). For instance, Koop et al. (1994) and Koop et al. (1995) introduced the Gibbs sampling method for posterior inference in SFMs. Van den Broeck et al. (1994) suggested the Bayesian inference into SFMs, which is also fundamental to the development of KOS97. Also the development by Fernandez et al. (1997) for the panel data problem and by Osiewalski and Steel (1998) for easier computations of Bayesian SFA.

In the further development of the Bayesian approach in SFA, Griffin and Steel (2004) proposed

a semi-parametric Bayesian framework with a non-parametrically defined distribution of inefficiency, Atkinson and Dorfman (2005) considered the presence of undesirable output. While more recently, Tsionas and Kumbhakar (2014) used Bayesian inference to gain a robust estimation of inefficiency in their four error component model, which is in a similar form to the KLH14 and CKMV model with transitory and persistent inefficiency terms as discussed in the previous section. Griffiths and Hajargasht (2016) proposed two models considering the correlations between the explanatory variables and the error terms to deal with the endogeneity problem in Bayesian SFA. Liu et al. (2017) introduced a Bayesian framework to estimate variants of the cross-sectional and time-varying stochastic frontier factor model of Kneip et al. (2012).

In the context of healthcare, Koop et al. (1997) and Griffin and Steel (2004) illustrated their proposed methods with a data set of hospitals in the US, regarding the cost efficiency. In later studies, Widmer (2015) analyzed 90 Swiss public hospitals with Bayesian inference of a SFM and a random parameter frontier model, and found gains of cost efficiency during 2004 to 2009.

Overall, in the context of SFA, comparing to the conventional MLE, the flexibility and other advantages of Bayesian inference lead to it having a great potential in the efficiency analysis of healthcare and hospitals.

8 Concluding Remarks

In this chapter, we very briefly introduced a group of representative models in the wide spectrum of SFMs and provided illustrations with corresponding code in R. We mainly followed the development routine of three advancements over the basic SFM, where the panel data framework, determinants of inefficiency and semi-parametric models were considered, respectively. In each routine, we may find the advantages of the further developed models are usually at the cost of some new limitations. Generally, a more advanced model is usually conducted with more sophisticated model specifications, leading to a possibly lower degree of freedom, more assumptions and higher computational intensity, etc. Hence, choosing a model that is suitable for the research scenario is

critical for an analysis. Such importance was also emphasized in the empirical illustration regarding the public hospitals in Queensland, Australia, where we may find the estimation of inefficiency by different models varies dramatically.

Obviously, there are many more variants of SFMs developed in the last several decades than those aforementioned, such as the spatial SFA and Bayesian SFMs that we briefly introduced. Moreover, the stochastic frontier paradigm is still expanding by incorporating new techniques. For more detailed discussions of more functional models and techniques, readers may refer to Kumbhakar et al. (2015); Sickles and Zelenyuk (2019); Nguyen et al. (2022), to mention a few.

Acknowledgments

References

Aigner, D., Lovell, C., and Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics*, 6(1):21–37.

Atkinson, S. E. and Dorfman, J. H. (2005). Bayesian measurement of productivity and efficiency in the presence of undesirable outputs: Crediting electric utilities for reducing air pollution. *Journal of Econometrics*, 126(2):445–468.

Battese, G. and Coelli, T. (1992). Frontier production functions, technical efficiency and panel data: With application to paddy farmers in India. *Journal of Productivity Analysis*, 3(1–2):153–169.

Battese, G. E. and Coelli, T. J. (1995). A model for technical inefficiency effects in a stochastic frontier production function for panel data. *Empirical Economics*, 20(2):325–332.

Belotti, F., Daidone, S., Ilardi, G., and Atella, V. (2013). Stochastic frontier analysis using Stata. *Stata Journal*, 13(4):718–758.

Besstremyannaya, G. (2011). Managerial performance and cost efficiency of Japanese local public hospitals: A latent class stochastic frontier model. *Health Economics*, 20(S1):19–34.

Bogetoft, P. and Otto, L. (2020). *Benchmarking with DEA and SFA*. R package version 0.29.

Caudill, S. B., Ford, J. M., and Gropper, D. M. (1995). Frontier estimation and firm-specific inefficiency measures in the presence of heteroscedasticity. *Journal of Business & Economic Statistics*, 13(1):105–111.

Chirikos, T. N. and Sear, A. M. (2000). Measuring hospital efficiency: A comparison of two approaches. *Health Services Research*, 34(6):1389.

Coelli, T. and Henningsen, A. (2020). *frontier: Stochastic Frontier Analysis*. R package version 1.1-8.

Colombi, R., Kumbhakar, S. C., Martini, G., and Vittadini, G. (2014). Closed-skew normality in stochastic frontiers with individual effects and long/short-run efficiency. *Journal of Productivity Analysis*, 42(2):123–136.

Colombi, R., Martini, G., and Vittadini, G. (2017). Determinants of transient and persistent hospital efficiency: The case of Italy. *Health Economics*, 26:5–22.

Comans, T., Nguyen, K.-H., Zelenyuk, V., Gray, L., Tran, A., Nguyen, B.-H., Wang, Z., and Moretto, N. (2020). The cost of residential aged care, Technical Report Research Paper 9, The Royal Commission into Aged Care Quality and Safety. Technical report.

Cornwell, C., Schmidt, P., and Sickles, R. C. (1990). Production frontiers with cross-sectional and time-series variation in efficiency levels. *Journal of Econometrics*, 46(1):185–200.

Croissant, Y. and Millo, G. (2008). Panel data econometrics in R: The plm package. *Journal of Statistical Software*, 27(2):1–43.

Dakpo, K., Desjeux, Y., and Latruffe, L. (2021). *sfaR: Stochastic Frontier Analysis using R*. R package version 0.1.0.

Daraio, C. and Simar, L. (2007). *Economies of scale, scope and experience in the Italian motor-vehicle sector*, chapter 6, pages 135–165. Springer.

Fan, Y., Li, Q., and Weersink, A. (1996). Semiparametric estimation of stochastic production frontier models. *Journal of Business & Economic Statistics*, 14(4):460–468.

Farsi, M. and Filippini, M. (2008). Effects of ownership, subsidization and teaching activities on hospital costs in Switzerland. *Health Economics*, 17(3):335–350.

Farsi, M., Filippini, M., and Kuenzle, M. (2005). Unobserved heterogeneity in stochastic cost frontier models: an application to Swiss nursing homes. *Applied Economics*, 37(18):2127–2141.

Fernandez, C., Osiewalski, J., and Steel, M. F. (1997). On the use of panel data in stochastic frontier models with improper priors. *Journal of Econometrics*, 79(1):169–193.

Filippini, M. and Greene, W. H. (2016). Persistent and transient productive inefficiency: A maximum simulated likelihood approach. *Journal of Productivity Analysis*, 45(2):187–196.

Filippini, M., Masiero, G., and Santarossa, M. (2021). Productivity change and efficiency in the Swiss nursing home industry. *Applied Economics*, pages 1–14.

Fuller, C. and Sickles, R. (2022). Homelessness on the West Coast: What role does health play? Technical report, Rice University.

Gabbitas, O. and Jeffs, C. (2009). Assessing productivity in the delivery of public hospital services in Australia: Some experimental estimates-Productivity Commission conference paper. Technical report, Productivity Commission.

Glass, A., Kenjegalieva, K., and Paez-Farrell, J. (2013). Productivity growth decomposition using a spatial autoregressive frontier model. *Economics Letters*, 119(3):291–295.

Glass, A., Kenjegalieva, K., and Sickles, R. C. (2014). Estimating efficiency spillovers with state level evidence for manufacturing in the US. *Economics Letters*, 123(2):154–159.

Glass, A., Kenjegalieva, K., Sickles, R. C., and Weyman-Jones, T. G. (2016a). The spatial efficiency multiplier and random effects in spatial stochastic frontier models. RISE Working Paper No. 16-002.

Glass, A. J., Kenjegalieva, K., and Sickles, R. (2012). The economic case for the spatial error model with an application to state vehicle usage in the US. *Rice University manuscript*.

Glass, A. J., Kenjegalieva, K., and Sickles, R. C. (2016b). A spatial autoregressive stochastic frontier model for panel data with asymmetric efficiency spillovers. *Journal of Econometrics*, 190(2):289–300.

Gong, B. H. and Sickles, R. C. (1992). A comparison between stochastic frontier and data envelopment methods using panel data. *Journal of Econometrics*, 51(1992):259–284.

Greene, W. H. (2005). Fixed and random effects in stochastic frontier models. *Journal of Productivity Analysis*, 23(1):7–32.

Griffin, J. E. and Steel, M. F. (2004). Semiparametric Bayesian inference for stochastic frontier models. *Journal of Econometrics*, 123(1):121–152.

Griffiths, W. E. and Hajargasht, G. (2016). Some models for stochastic frontiers with endogeneity. *Journal of Econometrics*, 190(2):341–348.

Grosskopf, S. and Valdmanis, V. (1987). Measuring hospital performance: A non-parametric approach. *Journal of Health Economics*, 6(2):89–107.

Hayfield, T. and Racine, J. S. (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software*, 27(5).

Herr, A. (2008). Cost and technical efficiency of German hospitals: does ownership matter? *Health Economics*, 17(9):1057–1071.

Hollingsworth, B. (2008). The measurement of efficiency and productivity of health care delivery. *Health Economics*, 17(10):1107–1128.

Jondrow, J., Lovell, C. K., Materov, I. S., and Schmidt, P. (1982). On the estimation of technical inefficiency in the stochastic frontier production function model. *Journal of Econometrics*, 19(2–3):233–238.

Kim, Y. and Schmidt, P. (2000). A review and empirical comparison of bayesian and classical approaches to inference on efficiency levels in stochastic frontier models with panel data. *Journal of Productivity Analysis*, 14(2):91–118.

Kneip, A., Sickles, R. C., and Song, W. (2012). A new panel data treatment for heterogeneity in time trends. *Econometric Theory*, 28(3):590–628.

Koop, G., Osiewalski, J., and Steel, M. F. (1994). Bayesian Efficiency Analysis with a Flexible Form: The AIM Cost Function. *Journal of Business & Economic Statistics*, 12(3):339–346.

Koop, G., Osiewalski, J., and Steel, M. F. (1997). Bayesian Efficiency Analysis Through Individual Effects: Hospital Cost Frontiers. *Journal of Econometrics*, 76(1):77–105.

Koop, G., Steel, M. F., and Osiewalski, J. (1995). Posterior analysis of stochastic frontier models using Gibbs sampling. *Computational Statistics*, 10:353–373.

Kumbhakar, S. C. (1990). Production frontiers, panel data, and time-varying technical inefficiency. *Journal of Econometrics*, 46(1):201–211.

Kumbhakar, S. C., Ghosh, S., and McGuckin, J. T. (1991). A generalized production frontier approach for estimating determinants of inefficiency in US dairy farms. *Journal of Business & Economic Statistics*, 9(3):279–286.

Kumbhakar, S. C., Lien, G., and Hardaker, J. B. (2014). Technical efficiency in competing panel data models: a study of Norwegian grain farming. *Journal of Productivity Analysis*, 41(2):321–337.

Kumbhakar, S. C., Park, B. U., Simar, L., and Tsionas, E. G. (2007). Nonparametric Stochastic Frontiers: A Local Maximum Likelihood Approach. *Journal of Econometrics*, 137(1):1–27.

Kumbhakar, S. C., Parmeter, C. F., and Zelenyuk, V. (2022a). *Stochastic frontier analysis: Foundations and advances I*. Springer, Singapore. forthcoming.

Kumbhakar, S. C., Parmeter, C. F., and Zelenyuk, V. (2022b). *Stochastic frontier analysis: Foundations and advances II*. Springer, Singapore. forthcoming.

Kumbhakar, S. C., Wang, H., and Horncastle, A. P. (2015). *A practitioner's guide to stochastic frontier analysis using Stata*. Cambridge University Press.

Kumbhakar, S. C. and Wang, H.-J. (2005). Estimation of growth convergence using a stochastic production frontier approach. *Economics Letters*, 88:300–305.

Lampe, H. W. and Hilgers, D. (2015). Trajectories of efficiency measurement: A bibliometric analysis of DEA and SFA. *European Journal of Operational Research*, 240(1):1–21.

Le Gallo, J., Baumont, C., Dall'erba, S., and Ertur, C. (2005). On the property of diffusion in the spatial error model. *Applied Economics Letters*, 12(9):533–536.

Li, T. and Rosenman, R. (2001). Cost inefficiency in Washington hospitals: A stochastic frontier approach using panel data. *Health Care Management Science*, 4(2):73–81.

Linna, M. (1998). Measuring hospital cost efficiency with panel data models. *Health Economics*, 7(5):415–427.

Liu, J., Sickles, R. C., and Tsionas, E. G. (2017). Bayesian treatments to stochastic panel data models with an application to models of productivity. Special Issue of *Econometrics* on Recent Developments in Panel Data Methods, Choi and Ryo Okui (Guest-Editors), 5:3:1–21.

Liu, W., Cheng, Q., Sickles, R. C., et al. (2022a). Productivity growth and spillovers across european industries: A global value chain perspective based on euro klems. Technical report, Rice University, mimeo.

Liu, W. and Sickles, R. C. (2021). Industry-specific productivity and spatial spillovers through input-output linkages. Technical report, Rice University, mimeo.

Liu, W., Sickles, R. C., and Zhao, Y. (2022b). Measuring productivity growth and technology spillovers through global value chains: Analysis of a US–Sino decoupling. In *Essays in Honor of M. Hashem Pesaran: Prediction and Macro Modeling*. Emerald Publishing Limited.

Meeusen, W. and van den Broeck, J. (1977). Efficiency estimation from Cobb-Douglas production functions with composed error. *International Economic Review*, 18(2):435–44.

Nghiem, S., Coelli, T., and Barber, S. (2011). Sources of productivity growth in health services: A case study of Queensland public hospitals. *Economic Analysis and Policy*, 41(1):37–48.

Nguyen, B. H., Sickles, R. C., and Zelenyuk, V. (2022). Efficiency Analysis with stochastic frontier models using popular statistical softwares. In Chotikapanich, D., Rambaldi, A., and Rohde,

N., editors, *Advances in Economic Measurement.* A Volume in Honour of D.S. Prasada Rao. Palgrave-Macmillan. Forthcoming.

Nguyen, B. H. and Zelenyuk, V. (2021a). Aggregate efficiency of industry and its groups: The case of Queensland public hospitals. *Empirical Economics*, 60(6):2795–2836.

Nguyen, B. H. and Zelenyuk, V. (2021b). Robust efficiency analysis of public hospitals in Queensland, Australia. In Daouia, A. and Ruiz-Gazen, A., editors, *Advances in Contemporary Statistics and Econometrics*, pages 221–242. Springer.

O'Donnell, C. and Nguyen, K. (2013). An econometric approach to estimating support prices and measures of productivity change in public hospitals. *Journal of Productivity Analysis*, 40(3):323–335.

Osiewalski, J. and Steel, M. F. (1998). Numerical Tools for the Bayesian Analysis of Stochastic Frontier Models. *Journal of Productivity Analysis*, 10(1):103–117.

Parmeter, C. F. and Zelenyuk, V. (2019). Combining the virtues of stochastic frontier and data envelopment analysis. *Operations Research*, 67(6):1628–1658.

Pitt, M. M. and Lee, L.-F. (1981). The measurement and sources of technical inefficiency in the Indonesian weaving industry. *Journal of Development Economics*, 9(1):43–64.

R Core Team (2022). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria.

Rosko, M. D. (1999). Impact of internal and external environmental pressures on hospital inefficiency. *Health Care Management Science*, 2(2):63–74.

Rosko, M. D. (2001). Cost efficiency of US hospitals: a stochastic frontier approach. *Health Economics*, 10(6):539–551.

Rosko, M. D. (2004). Performance of US teaching hospitals: A panel analysis of cost inefficiency. *Health Care Management Science*, 7(1):7–16.

Rosko, M. D. (2022). A Review of U.S. Hospital SF Articles. Working paper.

Schmidt, P. and Sickles, R. C. (1984). Production frontiers and panel data. *Journal of Business & Economic Statistics*, 2(4):367–374.

See, K., Grosskopf, S., Valdmanis, V., and Zelenyuk, V. (2021). What do we know from the vast literature on efficiency and productivity in healthcare? A Systematic Review and Bibliometric Analysis. Working paper.

Sickles, R. C., Song, W., and Zelenyuk, V. (2020). Econometric analysis of productivity: Theory and implementation in R. In Vinod, H. D. and Rao, C., editors, *Financial, Macro and Micro Econometrics Using R*, volume 42 of *Handbook of Statistics*, pages 267 – 297. Elsevier.

Sickles, R. C. and Zelenyuk, V. (2019). *Measurement of Productivity and Efficiency: Theory and Practice*. New York, NY: Cambridge University Press.

Simar, L., Van Keilegom, I., and Zelenyuk, V. (2017). Nonparametric least squares methods for stochastic frontier models. *Journal of Productivity Analysis*, 47(3):189–204.

Tsionas, E. G. and Kumbhakar, S. C. (2014). Firm heterogeneity, persistent and transient technical inefficiency: A generalized true random-effects model. *Journal of Applied Econometrics*, 29(1):110–132.

Van den Broeck, J., Koop, G., Osiewalski, J., and Steel, M. F. J. (1994). Stochastic frontier models: A Bayesian perspective. *Journal of Econometrics*, 61(2):273–303.

Varabyova, Y. and Schreyögg, J. (2013). International comparisons of the technical efficiency of the hospital sector: Panel data analysis of OECD countries using parametric and non-parametric approaches. *Health Policy*, 112(1-2):70–79.

Vitaliano, D. F. and Toren, M. (1994). Cost and efficiency in nursing homes: A stochastic frontier approach. *Journal of Health Economics*, 13(3):281–300.

Vitikainen, K., Linna, M., and Street, A. (2010). Substituting inpatient for outpatient care: What is the impact on hospital costs and efficiency? *The European Journal of Health Economics*, 11(4):395–404.

Wagstaff, A. (1989). Estimating efficiency in the hospital sector: a comparison of three statistical cost frontier models. *Applied Economics*, 21(5):659–672.

Wang, H.-J. and Schmidt, P. (2002). One-step and two-step estimation of the effects of exogenous variables on technical efficiency levels. *Journal of Productivity Analysis*, 18(2):129–144.

Wang, Z. and Zelenyuk, V. (2022). Random vs. explained inefficiency in SFA: The case of Queensland hospitals. Working paper, Centre for Efficiency and Productivity Analysis, School of Economics, University of Queensland.

Widmer, P. K. (2015). Does prospective payment increase hospital (in) efficiency? Evidence from the Swiss hospital sector. *The European Journal of Health Economics*, 16(4):407–419.

Zuckerman, S., Hadley, J., and Iezzoni, L. (1994). Measuring hospital efficiency with frontier cost functions. *Journal of Health Economics*, 13(3):255–280.

A   Full R code for the empirical illustrations

```r
1  rm(list=ls())
2  graphics.off()
3  # Data Process
       --------------------------------------------------------------
4  # Read sample data
5  data <- read.csv("QLD.csv")
6  # Convert to panel data
7  library(plm)
8  paneldata<- pdata.frame(data, c("HOSID","Yeardummy"))
9  # Define formulas for different models
10 form = lAggout ~ lBEDS + lAgglabours + lSUPP
11 formz = lAggout ~ lBEDS + lAgglabours + lSUPP | TEACH + Remote + Small
12 formt = lAggout ~ lBEDS + lAgglabours + lSUPP + factor(HOSID)
13 # SFMs
       ---------------------------------------------------------------------
14 library(frontier)
15 ######################### ALS77 #########################
16 attach(data)
17 # Model estimation
18 als77 <- sfa (form, data = data, ineffDecrease = T, truncNorm = F,
       timeEffect = F)
19 summary(als77)
20 # Individual inefficiency by JLMS
21   # Coefficients
22   als77coef <- coef(als77, which = "mle", extraPar = T)
23   # epsilon from fitted values
24   fals77 <- fitted(als77, asInData = T)
25   ei = lAggout - fals77
26   # E(ui|ei)
27   us2 = (als77coef[["sigmaU"]])^2
28   vs2 = (als77coef[["sigmaV"]])^2
29   sigmastar = sqrt((vs2*us2)/(vs2+us2))
30   ustari = (-us2*ei)/(vs2+us2)
31   uals77 = ((sigmastar*dnorm(ustari/sigmastar))/(pnorm(ustari/sigmastar)))
       +ustari
32   inals77 = 1 - exp(-uals77)
33   summary(inals77)
34
35 ######################### SS84 #########################
36 attach(paneldata)
37 # Fixed effects for ui
38 ss84 <- plm(form, data = paneldata,
39             model = "within", index = c("HOSID","Yeardummy"), effect = "
                 individual")
40 summary(ss84)
41   # ai and uihat
42   ai = as.numeric(unname(fixef(ss84)))
43   uss84 = max(ai) - ai
44   inss84 = 1 - exp(-uss84)
45   summary(inss84)
```

```r
46
47 ######################### PL81 #########################
48 attach(paneldata)
49 # Fixed effects
50 pl81f <- plm(form, data = paneldata,
51              model = "random", index = c("HOSID","Yeardummy"), effect = "
                    individual")
52 summary(pl81f)
53   # Extract initial values
54   sigmau2 = as.numeric(ercomp(pl81f)[["sigma2"]][2])
55   sigmav2 = as.numeric(ercomp(pl81f)[["sigma2"]][1])
56   sigmasq = sigmau2+sigmav2
57   gamma = sigmau2/sigmasq
58
59   init_para = c(as.numeric(pl81f$coefficients), sigmasq, gamma)
60 # MLE with initial values
61 pl81 <- sfa (form, data = paneldata, ineffDecrease = T, truncNorm = F,
      timeEffect =F,
62              startVal = init_para)
63 summary(pl81)
64
65   # Estimate individual inefficiency with user-written functions
66   inpl81 <- 1-efficiencies (pl81, asInData = T, logDepVar = T, minusU = T)
67   summary(inpl81)
68
69 ######################### CSS90 #########################
70 attach(paneldata)
71 # Fixed effects for ui
72 c90ori <- plm(form, data = paneldata,
73                model = "within", index = c("HOSID","Yeardummy"), effect =
                        "individual")
74 summary(c90ori)
75   # Vectors for regression with t
76   ai = as.numeric(c90ori[["residuals"]])
77   t = c(1, 2, 3, 4)
78   t2 = t^2
79   ait = numeric()
80 # Regression by unit
81 for (i in 1:length(unique(HOSID))){
82   aittemp = lm(ai[(4*(i-1)+1):(4*(i-1)+4)] ~ t2 + t)$fitted.values
83   aittemp = as.numeric(unname(aittemp))
84   ait = c(ait, aittemp)
85 }
86 c90 = cbind(ait, t)
87   # cthat for each year
88   ajtyear = aggregate(ait ~ t, data = c90, max)
89   # Compute uithat
90   c90 = as.data.frame(cbind(c90, ajtyear[,2]))
91   colnames(c90) <- c("ait","t","ajt")
92   uc90 = c90$ajt - c90$ait
93   summary(uc90)
94   inc90 = 1 - exp(-uc90)
95   summary(inc90)
96
```

```
97  ######################### BC92 #########################
98  attach(paneldata)
99  # Model estimation
100 bc92 <- sfa (form, data = paneldata, ineffDecrease = T, truncNorm = T,
        timeEffect = T)
101 summary(bc92, effic = F, logDepVar = T, effMinusU = T)
102
103   # Estimate individual inefficiency with user-written functions
104   inbc92 <- 1-efficiencies (bc92, asInData = T, logDepVar = T, minusU = T)
105   summary(inbc92)
106
107 ######################### G05 #########################
108 attach(paneldata)
109 # Model estimation
110 g05 <- sfa (formt, data = paneldata, ineffDecrease = T, truncNorm = F,
        timeEffect = T)
111 summary(g05, effic = F, logDepVar = T, effMinusU = T)
112
113   # Estimate individual inefficiency with user-written functions
114   ing05 <- 1-efficiencies (g05, asInData = T, logDepVar = T, minusU = T)
115   summary(ing05)
116
117 ######################### KLH14 #########################
118 attach(paneldata)
119 # Fixed effects for alphai and epsilonit
120 klh14ori <- plm(form, data = paneldata,
121                 model = "within", index = c("HOSID","Yeardummy"), effect = "
                        individual")
122 summary(klh14ori)
123 # alphai
124 alphai = as.numeric(unname(fixef(klh14ori)))
125 # epsilonit
126 epsiloni = as.numeric(klh14ori[["residuals"]])
127 summary(alphai)
128 summary(epsiloni)
129 # Constant for sfa estimation
130 constant = 1
131 # Persistent inefficiency
132 klh14per = as.data.frame(cbind(alphai,constant))
133 klh14p <- sfa(alphai~constant-1, data = klh14per, ineffDecrease = T,
        truncNorm = F, timeEffect = F)
134 summary(klh14p, effic = F, logDepVar = T, effMinusU = T)
135
136   # Estimate individual inefficiency with user-written functions
137   inklh14p <- 1-efficiencies (klh14p, asInData = T, logDepVar = T, minusU
        = T)
138   summary(inklh14p)
139
140 # Transitory inefficiency
141 klh14tran = as.data.frame(cbind(epsiloni,constant))
142 klh14t <- sfa(epsiloni~constant-1, data = klh14tran, ineffDecrease = T,
        truncNorm = F, timeEffect = T)
143 summary(klh14t, effic = F, logDepVar = T, effMinusU = T)
144
```

```r
145    # Estimate individual inefficiency with user-written functions
146    inklh14t <- 1-efficiencies (klh14t, asInData = T, logDepVar = T, minusU
           = T)
147    summary(inklh14t)
148
149 ######################### BC95 #########################
150 attach(paneldata)
151 # Model estimation
152 bc95 <- sfa(formz, data = paneldata, ineffDecrease = T, truncNorm = T,
        timeEffect = F)
153 summary(bc95, effic = F, logDepVar = T, effMinusU = T)
154
155    # Estimate individual inefficiency with user-written functions
156    inbc95 <- 1-efficiencies (bc95, asInData = T, logDepVar = T, minusU = T)
157    summary(inbc95)
158 ######################### SVKZ #########################
159 library(np)
160 # Define factor variables and formulas
161 fTEACH = as.factor(TEACH)
162 fRemote = as.factor(Remote)
163 fSmall = as.factor(Small)
164
165 forms = lAggout ~ lBEDS + lAgglabours + lSUPP + fTEACH + fRemote + fSmall
166 forms3 = ehat3 ~ lBEDS + lAgglabours + lSUPP + fTEACH + fRemote + fSmall
167
168 # Bandwidths selection
169 bws.r1 <- npregbw(forms, regtype="ll", data=data, bwmethod = "cv.ls",
        ckertype = "epanechnikov")
170 # Estimate conditional mean and extract fitted values and residuals
171 r1.est <- npreg(bws=bws.r1, gradients=TRUE)
172 r1hat    <- fitted(r1.est)
173 e1hat   <- residuals(r1.est)
174 # Generate for r3
175 ehat3 <- e1hat^3
176
177 # Bandwidths selection for r3 (skewness measures)
178 bws.r3 <- npregbw(forms3, data=data, regtype="ll", bwmethod = "cv.ls",
        ckertype = "epanechnikov")
179 # Estimate conditional mean and extract fitted values and residuals
180 r3.est  <- npreg(bws=bws.r3, gradients=TRUE)
181 r3hat  <- fitted(r3.est)
182 e3hat  <- residuals(r3.est)
183
184    # Estimate individual inefficiency
185    sigu3.hat <- sqrt(pi/2)*(pi/(pi-4))*r3hat
186    sigu.hat <- apply(cbind(0,sigu3.hat),1,FUN=max)^(1/3)
187    muhat <- sqrt(2/pi)*sigu.hat
188    summary(muhat)
189    insvkz = 1 - exp(-muhat)
190    summary(insvkz)
191
192 # Statistics of results
       ---------------------------------------------------
193    library(qpcR)
```

```r
194    # Combine inefficiency vectors of different length
195    ineffs <- qpcR:::cbind.na(inals77, inss84, inpl81, inc90, inbc92, ing05,
           inklh14t, inklh14p,inbc95, insvkz)
196    # Apply statistical analysis to each model
197    stat = list("mean" = apply(ineffs, 2, mean, na.rm = T),
198                "sd" = apply(ineffs, 2, sd, na.rm = T),
199                "min" = apply(ineffs, 2, min, na.rm = T),
200                "Q1" = apply(ineffs, 2, quantile, probs=0.25, na.rm = T),
201                "Median" = apply(ineffs, 2, quantile, probs=0.5, na.rm = T),
202                "Q3" = apply(ineffs, 2, quantile, probs=0.75, na.rm = T),
203                "max" = apply(ineffs, 2, max, na.rm = T))
204
205    write.csv(stat, file = 'Stats of inefficiency.csv')
206
207 # Kernel densities of estimations
       -------------------------------------------
208    require("ggplot2")
209    .df <- na.omit(data.frame(x = inals77))
210    .nbins <- pretty(range(.df$x), n = nclass.FD(.df$x), min.n = 1)
211    .als<-ggplot(data = .df, aes(x = x, y = ..density..)) +
212      # Epanechnikov kernel and CV bandwidth
213      geom_density(
214        kernel = "gaussian",
215        bw = "ucv",
216        alpha = 0.5,
217        #Here for single group: color and fill without aes()
218        color = "brown1", fill = "brown1",
219        #shut the legend
220        show.legend = FALSE
221      ) +
222      scale_y_continuous(expand = c(0.01, 0)) +
223      xlab("ALS77") +
224      ylab("Density") +
225
226      RcmdrPlugin.KMggplot2::theme_simple(base_size = 14, base_family = "
           sans")
227    print(.als)
228    rm(.df, .nbins)
229
230    require("ggplot2")
231    .df <- na.omit(data.frame(x = inss84))
232    .nbins <- pretty(range(.df$x), n = nclass.FD(.df$x), min.n = 1)
233    .ss84<-ggplot(data = .df, aes(x = x, y = ..density..)) +
234      # Epanechnikov kernel and CV bandwidth
235      geom_density(
236        kernel = "gaussian",
237        bw = "ucv",
238        alpha = 0.5,
239        #Here for single group: color and fill without aes()
240        color = "darkgreen", fill = "darkgreen",
241        #shut the legend
242        show.legend = FALSE
243      ) +
244      scale_y_continuous(expand = c(0.01, 0)) +
```

```
245      xlab("SS84") +
246      ylab("Density") +
247
248      RcmdrPlugin.KMggplot2::theme_simple(base_size = 14, base_family = "
             sans")
249  print(.ss84)
250  rm(.df, .nbins)
251
252  require("ggplot2")
253  .df <- na.omit(data.frame(x = inpl81))
254  .nbins <- pretty(range(.df$x), n = nclass.FD(.df$x), min.n = 1)
255  .pl81<-ggplot(data = .df, aes(x = x, y = ..density..)) +
256      # Epanechnikov kernel and CV bandwidth
257      geom_density(
258        kernel = "gaussian",
259        bw = "ucv",
260        alpha = 0.5,
261        #Here for single group: color and fill without aes()
262        color = "dodgerblue", fill = "dodgerblue",
263        #shut the legend
264        show.legend = FALSE
265      ) +
266      scale_y_continuous(expand = c(0.01, 0)) +
267      xlab("PL81") +
268      ylab("Density") +
269
270      RcmdrPlugin.KMggplot2::theme_simple(base_size = 14, base_family = "
             sans")
271  print(.pl81)
272  rm(.df, .nbins)
273
274  require("ggplot2")
275  .df <- na.omit(data.frame(x = inc90))
276  .nbins <- pretty(range(.df$x), n = nclass.FD(.df$x), min.n = 1)
277  .c90<-ggplot(data = .df, aes(x = x, y = ..density..)) +
278      # Epanechnikov kernel and CV bandwidth
279      geom_density(
280        kernel = "gaussian",
281        bw = "ucv",
282        alpha = 0.5,
283        #Here for single group: color and fill without aes()
284        color = "gold", fill = "gold",
285        #shut the legend
286        show.legend = FALSE
287      ) +
288      scale_y_continuous(expand = c(0.01, 0)) +
289      xlab("C90") +
290      ylab("Density") +
291
292      RcmdrPlugin.KMggplot2::theme_simple(base_size = 14, base_family = "
             sans")
293  print(.c90)
294  rm(.df, .nbins)
295
```

```
296    require("ggplot2")
297    .df <- na.omit(data.frame(x = inbc92))
298    .nbins <- pretty(range(.df$x), n = nclass.FD(.df$x), min.n = 1)
299    .bc92<-ggplot(data = .df, aes(x = x, y = ..density..)) +
300      # Epanechnikov kernel and CV bandwidth
301      geom_density(
302        kernel = "gaussian",
303        bw = "ucv",
304        alpha = 0.5,
305        #Here for single group: color and fill without aes()
306        color = "blueviolet", fill = "blueviolet",
307        #shut the legend
308        show.legend = FALSE
309      ) +
310      scale_y_continuous(expand = c(0.01, 0)) +
311      xlab("BC92") +
312      ylab("Density") +
313
314      RcmdrPlugin.KMggplot2::theme_simple(base_size = 14, base_family = "
             sans")
315    print(.bc92)
316    rm(.df, .nbins)
317
318    require("ggplot2")
319    .df <- na.omit(data.frame(x = ing05))
320    .nbins <- pretty(range(.df$x), n = nclass.FD(.df$x), min.n = 1)
321    .g05<-ggplot(data = .df, aes(x = x, y = ..density..)) +
322      # Epanechnikov kernel and CV bandwidth
323      geom_density(
324        kernel = "gaussian",
325        bw = "ucv",
326        alpha = 0.5,
327        #Here for single group: color and fill without aes()
328        color = "coral1", fill = "coral1",
329        #shut the legend
330        show.legend = FALSE
331      ) +
332      scale_y_continuous(expand = c(0.01, 0)) +
333      xlab("G05") +
334      ylab("Density") +
335
336      RcmdrPlugin.KMggplot2::theme_simple(base_size = 14, base_family = "
             sans")
337    print(.g05)
338    rm(.df, .nbins)
339
340    require("ggplot2")
341    # Disable scientific notation
342    options(scipen=1000)
343
344    .df <- na.omit(data.frame(x = inklh14t))
345    .nbins <- pretty(range(.df$x), n = nclass.FD(.df$x), min.n = 1)
346    .klh14t<-ggplot(data = .df, aes(x = x, y = ..density..)) +
347      # Epanechnikov kernel and CV bandwidth
```

```
348     geom_density(
349       kernel = "gaussian",
350       bw = "ucv",
351       alpha = 0.5,
352       #Here for single group: color and fill without aes()
353       color = "darkolivegreen1", fill = "darkolivegreen1",
354       #shut the legend
355       show.legend = FALSE
356     ) +
357     scale_y_continuous(expand = c(0.01, 0)) +
358     xlab("KLH14-Transitory") +
359     ylab("Density") +
360
361     RcmdrPlugin.KMggplot2::theme_simple(base_size = 14, base_family = "
              sans")
362   print(.klh14t)
363   rm(.df, .nbins)
364
365   require("ggplot2")
366   .df <- na.omit(data.frame(x = inklh14p))
367   .nbins <- pretty(range(.df$x), n = nclass.FD(.df$x), min.n = 1)
368   .klh14p<-ggplot(data = .df, aes(x = x, y = ..density..)) +
369     # Epanechnikov kernel and CV bandwidth
370     geom_density(
371       kernel = "gaussian",
372       bw = "ucv",
373       alpha = 0.5,
374       #Here for single group: color and fill without aes()
375       color = "royalblue", fill = "royalblue",
376       #shut the legend
377       show.legend = FALSE
378     ) +
379     scale_y_continuous(expand = c(0.01, 0)) +
380     xlab("KLH14-Persistent") +
381     ylab("Density") +
382
383     RcmdrPlugin.KMggplot2::theme_simple(base_size = 14, base_family = "
              sans")
384   print(.klh14p)
385   rm(.df, .nbins)
386
387   require("ggplot2")
388   .df <- na.omit(data.frame(x = inbc95))
389   .nbins <- pretty(range(.df$x), n = nclass.FD(.df$x), min.n = 1)
390   .bc95<-ggplot(data = .df, aes(x = x, y = ..density..)) +
391     # Epanechnikov kernel and CV bandwidth
392     geom_density(
393       kernel = "gaussian",
394       bw = "ucv",
395       alpha = 0.5,
396       #Here for single group: color and fill without aes()
397       color = "khaki3", fill = "khaki3",
398       #shut the legend
399       show.legend = FALSE
```

```r
400      ) +
401      scale_y_continuous(expand = c(0.01, 0)) +
402      xlab("BC95") +
403      ylab("Density") +
404
405      RcmdrPlugin.KMggplot2::theme_simple(base_size = 14, base_family = "
            sans")
406    print(.bc95)
407    rm(.df, .nbins)
408
409    require("ggplot2")
410    .df <- na.omit(data.frame(x = insvkz))
411    .nbins <- pretty(range(.df$x), n = nclass.FD(.df$x), min.n = 1)
412    .svkz<-ggplot(data = .df, aes(x = x, y = ..density..)) +
413      # Epanechnikov kernel and CV bandwidth
414      geom_density(
415        kernel = "gaussian",
416        bw = "ucv",
417        alpha = 0.5,
418        #Here for single group: color and fill without aes()
419        color = "mediumorchid1", fill = "mediumorchid1",
420        #shut the legend
421        show.legend = FALSE
422      ) +
423      scale_y_continuous(expand = c(0.01, 0)) +
424      xlab("SVKZ") +
425      ylab("Density") +
426
427      RcmdrPlugin.KMggplot2::theme_simple(base_size = 14, base_family = "
            sans")
428    print(.svkz)
429    rm(.df, .nbins)
430
431    library("ggpubr")
432    Inefficiencies <- ggarrange(.als, .ss84, .pl81, .c90, .bc92, .g05, .
         klh14t, .klh14p, .bc95, .svkz,
433                               ncol = 2, nrow = 5)
434    Inefficiencies
435 #--------------------END----------------------------
```