Journal of Productivity Analysis (2026) 65:6
https://doi.org/10.1007/s11123-025-00786-y

Incorporating intermediate outputs into SFA using an equation system approach with application to transport

VHT Tran[1] · P. Wheat[1] · AD Stead[1] · ASJ Smith[1]

Received: 28 April 2024 / Accepted: 28 October 2025

Abstract
Conventional Stochastic Frontier Analysis conceptualizes a firm's production process as a single stage of transforming inputs to final outputs. In practice many firms have a multi-stage production process and conceptualizing this as a single stage only assesses overall efficiency and does not provide insight into where inefficiency resides which is of greater interest to decision making units and policy makers. The paper addresses this issue by developing a Network Stochastic Frontier model into which intermediate output is incorporated. We propose a Network Stochastic Frontier model in which multiple stages of a production process are modelled as a system of equations. However, departing from Huang et al. (2017, 2018), our model focuses on how stage efficiency can be aggregated to measure overall efficiency and how the model exploits different data aggregations to yield more information about the performance of firms. Besides the model development, we make a contribution to the estimation method by introducing a simplified multi-step estimation procedure with an authors' written package in Stata, "networkSFA". The applicability of the model and the validity of the estimation method are demonstrated by an empirical example of English road maintenance.

Keywords Multi-stage process · Network SFA · Stage efficiency · Overall efficiency · Intermediate outputs

JEL D24 · L51 · R42

1 Introduction

Frontier analysis is commonly used to evaluate performance of decision making units (DMUs). It implicitly assumes that the production process of DMUs can be represented as a single-stage in which inputs are directly transformed to final outputs and there is no need to consider the role of intermediate outputs involved in the production process. This approach has the drawback that it provides limited insight into the performance of DMUs which produce and use intermediate outputs, and reveals nothing of the extent to which the overall inefficiency of the DMU consists of inefficiency in the production or in the later use of these (potentially endogenous) intermediate outputs. As discussed in several received papers in the efficiency literature (e.g. Charnes et al. 1986; Lin and Hong 2020), the production process of many DMUs can be broken down into at least two stages. The first stage produces intermediate outputs which are used to produce final outputs in the second stage (Fig. 1).[1]

Infrastructure industries provide many examples of multi-stage production processes. These are often subject to economic regulation which in turn involves using frontier analysis to assess the scope for efficiency savings. Infrastructure managers will undertake some maintenance and reconstruction of the infrastructure (intermediate outputs) to improve assets' quality (final outputs). Thus, the firm needs to transform inputs into intermediate outputs in an efficient manner but also choose and target the intermediate outputs to maximize final outputs. For example, the production process of highway managers who are responsible

✉ VHT Tran
V.T.Tran@leeds.ac.uk

[1] Institute for Transport Studies (ITS), University of Leeds, Leeds, UK

[1] We recognize that "multi-stages" could be interpreted differently in the literature. In this paper, multi-stages are related to two key decisions of DMUs on how they produce outputs. The first is how to produce intermediate output and the second is how intermediate output is transferred into final output. Thus, it does not imply time-lags between production stages, i.e. the production of intermediate goods and final goods could be a simultaneous process.

Fig. 1 Simple two-stage production process (where DMUs maximize final outputs)

for maintaining the usability of roads can be seen as a two-stage process. In the first stage, Local Highway Authorities (LHAs) implement treatments such as overlaying, resurfacing and patching to produce an intermediate output (treated road kilometers). These treatments can be procured in differing ways, yielding different efficiency outcomes even at this intermediate stage. In the second stage, the intermediate output contributes to improvements in overall road condition, which serves as the final output.[2] Any inappropriate choice in treatment types and locations receiving treatment is likely to cause inefficiency in the second stage.

Motivated by the need to measure efficiency of DMUs at stages (stage efficiency from now on), several authors have attempted to modify conventional efficiency models of Data Envelopment Analysis (DEA) toward the multi-stage approach, Network DEA (Chiou et al. 2010). The simplest Network DEA is independent DEA that applies conventional DEA to each stage in an isolated manner. The literature has expanded to more holistic analysis such as relational DEA which considers dependence between stages and allows estimation of stage efficiency along with overall efficiency. Despite the growth of DEA in this matter, there is a lack of attention to Network SFA (Peyrache and Silva, 2022). To the best of the authors' knowledge, only Huang et al. (2017, 2018) sought to develop an economic model equivalent to Network DEA, which they called copula-based Network SFA. However, their copula-based estimation is not straightforward. Thus, it might be not easy to implement their estimation method due to its computational complexity. In addition, the authors focus on predicting stage efficiency and do not compute overall efficiency. This underscores the importance of advancing Network SFA to bring it in line with the methodological sophistication of Network DEA.

This study is motivated by the potential to gain new insights into the efficiency of infrastructure industries where data on intermediate outputs are observable. Currently, such data are rarely incorporated in SFA yet they are critical for better estimating overall inefficiency and for pinpointing its source – whether inefficiency is more pronounced in the production of intermediate outputs, or in their transformation into final outputs. To address this, we develop a Network SF approach that exploits the intermediate output data to provide a more nuanced analysis of the performance of DMUs.

Our Network SF model offers several advantages over conventional (single-stage) SF model. First, by taking account of the multi-stage nature of production, our model provides indicators of not only overall efficiency but also stage efficiencies. Since overall efficiency is calculated as an aggregate of stage efficiencies, our model avoids overestimating overall efficiency. Second, the model yields a more granular understanding of the production process by estimating returns to scale at each stage that can be aggregated to derive the overall return to scale for the entire production process. Finally, the model allows incorporating different levels of data aggregation to assess the performance of not only DMUs but also their divisions.[3] This property is useful since we often see in reality that data of costs and inputs are aggregated at the DMU level due to un-allocable central costs and shared inputs, but the data of (intermediate and final) outputs are available at a division level. (See Subsection 4.2. and Appendix 2 for detailed discussion).

We differentiate our paper from Huang et al. (2017, 2018) in three aspects. First, we will build on the Network SF model of Huang et al. (2017, 2018) to enable evaluating overall efficiency (along with stage efficiency) and exploiting different data aggregation. Second, we propose a different multi-step estimation procedure (with the authors' written Stata package) which is simple to apply for practitioners. Monte Carlo simulations suggest that our simple estimation method is as good as that of Huang et al. (2017, 2018). Finally, we specify our model under different assumptions of endogeneity and DMUs' behaviors, and outline how our estimation method can address the endogeneity (if it exists). These were not addressed in Huang et al. (2017, 2018). To demonstrate the applicability of the model, we apply it to a two-stage road maintenance process of English LHAs. We then compare our model results with those of a single-stage model.

The remainder of this paper is structured as follows. Section 2 provides a literature review on the development of Network DEA and Network SFA. Section 3 discusses the contribution of this study. Section 4 consists of two subsections. Subsection 4.1 specifies our Network SF model followed by Subsection 4.2 discussing its benefits. Section 5 proposes a multi-step estimation method whose soundness is demonstrated by Monte Carlo simulations in Section 6. Section 7 applies the model to evaluate the performance

[2] The choice of road condition as a (final) output is often observed in the road maintenance literature such as Rouse and Chiu (2009), Ozbek et al. (2012) and Fallah-Fini et al. (2015), since it reflects the ultimate objective of maintenance activities which is to ensure the safety of the road network for use. Furthermore, it can be easily measurable via standardized metrics such as a roughness index and a pavement condition index.

[3] In this paper, "divisions" have a broad meaning. They refer to things by which the production process can be segmented such as different factories, activities, assets or regions. For instance, in our empirical example of road maintenance, different road types are considered as divisions.

of English LHAs in road maintenance with some reflection on what is the appropriate endogeneity and exogeneity assumption for regulated firms at the beginning of the section. Section 8 concludes.

2 Literature review

DEA and SFA are both widely used to measure efficiency. They evaluate efficiency of a DMU as its distance from an efficient frontier. The main difference between the two approaches is that the former is a nonparametric technique while the latter is a parametric one. Despite their relatively equal popularity in the general literature in efficiency, only applications of DEA have been established in multi-stage production processes thanks to its evolution in this matter over the recent decades. The development of SFA with respect to this issue seems to be limited.

Motivated by the need to evaluate efficiency arising at each stage, several models of Network DEA have been developed to take account of the multi-stage nature of the production process. According to Halkos et al (2014), these models can be classified into three main categories: independent DEA, connected DEA and relational DEA.[4]

Independent DEA represents each stage as a separate efficiency problem. It assumes that there is no connection between the stages and uses independent conventional DEA to estimate stage efficiency. Due to separate estimation, the approach could yield contradictory results and suggest conflicting strategies to improve performance (Chiou et al. 2010). Wang et al. (1997) and Seiford and Zhu (1999) are the first papers using independent DEA.

Unlike independent DEA, connected and relational DEA take account of dependence between the stages. Thus, they avoid conflicting findings of independent DEA and guarantee that DMUs are efficient overall only if they are efficient at all stages. The main difference between the two approaches is that the latter requires multipliers of intermediate outputs to be the same between two stages (so a mathematical link between overall efficiency and stage efficiency exists) while the former does not. The initial paper in Network DEA, Färe and Grosskopf (1996) is in the connected DEA category.

The first paper in relational DEA is Kao and Hwang (2008). The authors develop a multiplicative model where overall efficiency is a product of stage efficiency under constant returns to scale (CRS). Chen et al. (2009) extend Kao and Hwang's model to allow variable returns to scale. Under this model, overall efficiency is a weighted sum of its stage elements (an additive model).[5] It is noted that both models use a decomposition approach, i.e. overall efficiency is estimated and then decomposed to obtain stage efficiency. Recently, Despotis et al. (2016a, b) suggest a composition approach where stage efficiency is calculated first and then aggregated to estimate overall efficiency.

Network DEA has been developed to allow for the estimation of efficiency under various network structures (See Kao, 2014 for detail). As a result, Network DEA has been applied to many industries including transport services (Yu and Lin 2008; Tavassoli et al. 2014; Liu 2016; Lin and Hong 2020), insurance (Kao and Hwang 2008; Chen et al. 2009) and banking (Matthews 2013; Fukuyama and Matousek 2017).

In contrast to DEA, the application of SFA to multi-stage production problems is relatively underdeveloped due to the challenges of endogeneity and computational difficulty. To the best of the authors' knowledge, only Huang et al. (2017, 2018) attempted to develop an econometric model which is equivalent to Network DEA to estimate stage efficiency. The authors focused on shared inputs between the two production stages of banks. In the first stage, a proportion of labor and capital (inputs) is used to create deposits (an intermediate output). The intermediate output, along with the rest of the inputs are then employed as inputs in the second stage to produce loans, investments, and non-interest services (final outputs). The implication of a shared input is that the decision making in the two stages is not independent. Thus, stochastic frontier models of the stages have to be jointly estimated.

Huang et al. (2017.2018) suggested using a two-step estimation procedure to estimate their Network SF model. The first step applies a system of seemingly unrelated regressions (SUR) which takes account of correlation of composite errors between the regressions to simultaneously determine the proportional and slope parameters. The second step employs a copula-based maximum likelihood estimation (MLE) to estimate the intercepts of the regressions and the distributional parameters of the composite errors. Although the SUR estimation is relatively straightforward and easy to imitate by practitioners through use of existing estimation routines, the copula-based maximum likelihood estimation is not commonly available in statistical software. Thus, from the application perspective, it might be beneficial to develop a simplified estimation method. Furthermore, as Huang et al. (2017, 2018) aimed to estimate the share of inputs used in each stage and stage efficiency, they

[4] Also see Cook et al. (2010) and Kao (2014) for the literature reviews on Network DEA.

[5] In the output-oriented DEA of Chen et al. (2009), the weight is a share of outputs produced in each stage. In the input-oriented model, the weight is a share of inputs used in each stage.

neglected to measure overall efficiency nor discussed endogeneity in Network SFA (and in single-stage SFA).[6]

Before Huang et al. (2017, 2018), several papers attempted to use independent SFA to distinguish efficiency between stages. For example, Lan and Lin (2006) and Sjögren and Söderberg (2011) employed independent SFA to assess stage efficiency of railways and airline carriers. Similar to independent DEA, independent SFA fails to take account of the inter-stage dependence. This failure may lead to biased estimates of efficiency (Lai and Huang, 2013).[7] Furthermore, it reduces precision associated with the parameter estimates (Amsler et al. 2014). Like Huang et al. (2017, 2018), the papers neither measured overall efficiency using Network SFA nor discussed endogeneity.[8]

3 Contribution

Given the limited literature exploring this important problem in SFA, both in terms of model formulation and estimation method, we aim to build Network SF model that can be specified under different assumptions of exogeneity and DMUs' behaviors, and propose an estimator that seamlessly handles endogeneity while maintaining empirical tractability. The estimation method is deliberately simplified and an author written Stata package is provided with the aim of promoting further applications.

We distinguish our paper from Huang et al. (2017, 2018) in three main respects.

Firstly, we focus on providing measures of overall efficiency and different aggregations of data at different stages, rather than incorporating shared inputs across stages (which is not relevant in our framework). Our Network SF model allows aggregated data to be used in one stage but more disaggregated data in the other stage (See Subsection 3.2 and Appendix 2 for details). Therefore, the model better utilizes the available data, especially in the case of the road maintenance dataset to analyze performance at a DMU level and a division level (where possible). Hence this is expected to provide clear strategies on how inefficient DMUs can improve their performance. Furthermore, we explicitly show how overall inefficiency can be calculated by aggregating stage inefficiencies.

Secondly, for the estimation method, we will propose a relatively simple multi-step procedure which can be applied by practitioners. In the first step, we will yield estimates of slope parameters using SUR or three-stage least squares (3SLS), dependent on model specification. We will simplify the second estimation step of Huang et al. (2017, 2018) by using the method of moments (MM) (instead of the copula-based maximum likelihood estimation) to calculate distributional parameters and correct intercepts. The last step is to gauge efficiency. Our estimation method is simpler but not inferior to that of Huang et al (2017, 2018).

Thirdly, to make the model applicable in various cases, we will specify our Network SF model under different assumptions about DMUs' optimal behaviors and exogeneity. We will then discuss if the endogeneity appears in the Network model (and single-stage model) and how our estimation method can deal with it. This is not considered by Huang et al. (2017, 2018).

4 A Network stochastic frontier model

In this section, we introduce our Network SF approach under different assumptions about exogeneity and DMU behavior outlined in sub-section 4.1. We then discuss the advantages of our model over the conventional single-stage SFA in sub-section 4.2.

4.1 Model specification

Departing from the single-stage SF model, our Network SF model is a system of equations, each of which represents the production process of a particular stage of a production process.[9] Thus, stage-specific inefficiency, represented in the model as the non-negative error component of the equation relating to a particular stage of production, can be estimated. Overall inefficiency is then gauged by aggregating stage-specific inefficiencies. Thus, our model is comparable to the composition approach of Network DEA in this respect.

In the case of constant returns to scale at all stages of production, the overall inefficiency term is simply a sum of stage-specific inefficiency terms. However, this is not the case more generally, as we show below. For instance, in the case of exogenous final output (**case 1**), overall inefficiency is the first-stage inefficiency term plus the second-stage inefficiency term scaled by the intermediate output elasticity

[6] The authors did estimate efficiency of a single stage of production process where banks either use all labour and capital to produce deposit or use labour, capital and deposit to produce final outputs - loans, investments, and non-interest services. But they did not estimate overall efficiency using their Network SFA.

[7] Using a Monte Carlo simulation, Lai and Huang (2013) show that the estimates of variance of inefficiency and noise (i.e. $\hat{\sigma}_u^2$ and $\hat{\sigma}_v^2$) are much more biased if the dependence in the composite errors between equations of the system is not considered. As these biased variance estimates are used to predict efficiency, it distorts the results.

[8] Similar to Huang et al, (2017, 2018), Sjögren and Söderberg (2011) measured overall efficiency using single-stage SFA (not Network SFA).

[9] If the model is extended to exploit different data aggregation, one stage can be represented by multi equations (see subsection 3.2 and Appendix 2 for details).

from the first stage, represented in Eq. (4). Therefore, unless constant returns to scale exist in the first stage, overall efficiency is not the product of stage-specific efficiencies. Since our model allows unequal contribution of stage efficiency to overall efficiency (through a scale factor), our model is more in line with Network DEA of Chen et al (2009) rather than that of Kao and Hwang (2008).

In the interest of simplicity, but without loss of generality, we start with an assumption of a two-stage production process with only one input, one intermediate output and one final output.[10] We also assume that there are no shared inputs between the two stages, i.e. all inputs are used in the first stage and no inputs (rather than intermediate output) are involved in the second stage.[11] This is reasonable where we conceptualize intermediate outputs as treatments to infrastructure yielding higher or lower quality of infrastructure (final output). Although there are no shared inputs between the stages, the two stages are still dependent due to the presence of intermediate output at both stages (i.e. intermediate output is an output of the first stage but an input to the second stage). It is especially true for our empirical example of road maintenance where the two stages are a simultaneous process, thus the choice of treatment types and locations receiving treatments for example impact efficiency of both stages. This implies the dependency between the composite errors of the stages in our models.

We now specify our Network SF model under two conventional assumptions about DMUs' objective functions, but re-stated them for the multi-stage production process with the presence of intermediate outputs.[12] In addition to these established assumptions, we introduce a third case: the intermediate output is exogenous while the input and final output are endogenous. This assumption has not been considered under the single-stage production framework. Our model specification and the challenges in estimating overall efficiency in this case are discussed in Appendix 1.

Case 1: Final output is exogenous (input and intermediate output are endogenous). We make an additional assumption about the orientation of efficiency that DMUs pursue input-oriented efficiency at both stages.[13] This means that given a target final output, DMUs try to minimize intermediate output (in the second stage). They then minimize input in the first stage (given the intermediate output determined in the second stage). The model implies that DMUs know the amount of intermediate output needed to produce the target final output in the second stage.

Our Network SF model is a system of two equations specified as[14]

Stage 1:

$$\ln\left(x_i\right) = \alpha_1 + \beta_1 \ln\left(z_i\right) + u_{1i} + v_{1i} \tag{1}$$

Stage 2:

$$\ln\left(z_i\right) = \alpha_2 + \beta_2 \ln\left(y_i\right) + u_{2i} + v_{2i} \tag{2}$$

Where i denotes the specific DMU, x is the input, y is the final output, z is the intermediate output, u_1 and u_2 are inefficiency, v_1 and v_2 are error terms in the first and second stages respectively, the composite errors $\varepsilon_1 = u_1 + v_1$ and $\varepsilon_2 = u_2 + v_2$ are assumed to be correlated due to the dependency between the stages.

The equivalent conventional single-stage SF model (the reduced form of this system) is an input-oriented SF model created by substituting $\ln\left(z_i\right)$ in Eq. (1) with the right hand side of Eq. (2). The input-oriented single-stage SF model is

$$\begin{aligned} \ln\left(x_i\right) \quad = \quad & \left(\alpha_1 + \alpha_2\beta_1\right) \\ & + \beta_1\beta_2 \ln\left(y_i\right) + \left(u_{1i} + \beta_1 u_{2i}\right) + \left(v_{1i} + \beta_1 v_{2i}\right) \end{aligned} \tag{3}$$

Thus, (input-oriented) overall efficiency is

$$Exp(-u_i) = Exp\left(-\left(u_{1i} + \beta_1 u_{2i}\right)\right) \tag{4}$$

Case 2: The input is exogenous (the intermediate and final output are endogenous). We also assume that output-oriented efficiency is pursued in both stages; i.e. given fixed inputs, DMUs try to maximize intermediate output in the first stage, then maximize final output in the second stage.[15]

The Network SF model is a system of two equations specified as

[10] The model can be extended to a J-stage production process. See Appendix 3 for details.

[11] This assumption makes our model depart from that of Huang et al. (2017, 2018). It can be relaxed to allow additional independent inputs appearing in the second stage for cases 2 and 3.

[12] We decide not to combine the first two (conventional) assumptions to make the paper easier to follow as the model specification and the formula to calculate overall efficiency can be expressed explicitly for each case.

[13] For the equation systems specified in Eqs. (1) – (2) and Eqs. (5) – (6) input-oriented efficiency can be derived from output-oriented efficiency and vice versa. Thus the assumptions about the orientation of efficiency can be changed without significant impact on the estimation methods. However, for more complex models such as a system of non-homogeneous production functions, different estimation method is likely to be required if a different orientation of efficiency is assumed. See Kumbhakar and Tsionas (2006) for detail discussion.

[14] Note that (the log of) the intermediate outputs are in a linear form in both stage equations, so that overall efficiency is a function of stage efficiency and the scale – beta; no other variables impact overall efficiency. This holds for all the three cases.

[15] An example of output orientation is when DMUs attempt to maximize their profit given their fixed inputs in the first stage and then maximize their market value in the second stage.

Stage 1:

$$\ln(z_i) = \alpha_1 + \beta_1 \ln(x_i) - u_{1i} + v_{1i} \tag{5}$$

Stage 2:

$$\ln(y_i) = \alpha_2 + \beta_2 \ln(z_i) - u_{2i} + v_{2i} \tag{6}$$

The equivalent output-oriented single-stage SF model is

$$\begin{aligned} \ln(y_i) \quad = \quad & (\alpha_1\beta_2 + \alpha_2) + \beta_1\beta_2 \ln(x_i) \\ & -(\beta_2 u_{1i} + u_{2i}) + (\beta_2 v_{1i} + v_{2i}) \end{aligned} \tag{7}$$

Thus, (output-oriented) overall efficiency is

$$Exp(-u_i) = Exp\left(-\left(\beta_2 u_{1i} + u_{2i}\right)\right) \tag{8}$$

Table 1 summaries the endogeneity present in each case and model specified in this subsection.

Table 1 The endogeneity of the models in each case

Behavioral assumption	Two-stage SFA	Single-stage input oriented SFA	Single-stage output oriented SFA
Case 1: Final output is exogenous	Yes	No	Yes
Case 2: Input is exogenous	Yes	Yes	No

"Yes" means the endogeneity present in the model. "No" means there is no endogeneity in the model

4.2 Benefits of Network SFA

By taking account of the presence of the intermediate output in the production process and explicitly modelling each stage of production, our Network SF model offers several benefits (relative to the single-stage SFA).

First, our Network SF model improves performance indicators by avoiding over-estimated overall efficiency and providing additional information, i.e. stage efficiency.

To understand why estimating the model in closed form – the single-stage approach – may result in overestimation of DMU efficiency, consider the fact that the overall error term in the reduced form is obtained as a linear combination of the error terms of each of the individual equations. This summation will tend to dilute the skewness of the composed error. For example in Case 2, assuming independence of error terms, if we have J intermediate outputs, we have

$$\text{Skew}[\varepsilon_i] = \frac{\sum_{j=1}^{J} \beta_j^3 \text{Skew}[\varepsilon_{ij}] \left(\text{Var}[\varepsilon_{ij}]\right)^{3/2}}{\left(\sum_{j=1}^{J} \beta_j^2 \text{Skew}[\varepsilon_{ij}] \, \text{Var}[\varepsilon_{ij}]\right)^{3/2}}, \beta_J = 1.$$

Providing that no single variance term dominates the sum, the numerator grows roughly linearly with J, while the denominator scales as $J^{\frac{3}{2}}$. As a result, the skewness decays at a rate proportional to $J^{-\frac{1}{2}}$ (it is easiest to see this if we assume the error terms are identically distributed across equations), attenuating toward zero as the number of equations increases. Since identification of the inefficiency term in many SF specifications hinges on the skewness of the model residuals, this attenuation will tend to lead to underestimation of overall inefficiency, and increase the frequency of 'wrong skewness' problems. Estimating the structural equations, on the other hand, preserves the stronger skewness signal from each individual composed error, aiding identification. The overestimated overall efficiency of the single-stage SFA is found in our empirical example in Section 7.[16]

Furthermore, the single-stage approach only yields overall efficiency as a result of ignoring the multi-stage nature of the production. Although it is a good proxy for DMUs' performance, overall efficiency provides little information on how the performance can be improved since it does not reveal where inefficiency resides - whether inefficiency arises because of failure to procure (or deliver) intermediate outputs using the least possible inputs or costs versus whether sub-optimal mixes of intermediate outputs have been delivered to produce final outputs. The Network SF model addresses this issue by incorporating intermediate output into the model to provide insight into how DMUs perform in each stage - stage efficiency. This can be used to focus improvement efforts by the DMUs.

Second, our model provides more information about the production process.

By explicitly modelling the production process of each stage, our Network SF model provides insights into how inputs are transformed into intermediate outputs in the first stage and how final outputs is produced from these intermediate outputs in the second stage (i.e. economies of/returns to scale in each stage). This is impossible to analyze using the single stage model.

Finally, Network SFA allows for the incorporation of data at different levels of aggregation, and thus enhancing its applicability to diverse empirical settings.

Since the model is a system of equations we can exploit disaggregated data on one part of the production process should this be available. This subsection is to provide examples of how the Network SF model specified in Subsection

[16] A similar finding was found when we estimated the overall efficiency of rail-track renewals using the single-stage and Network SFA. Kao and Hwang (2008) provided an example of overestimating overall efficiency when the multi-stage process is ignored under the DEA framework.

4.1 can be extended to exploit different levels of data aggregation.

In reality, we often observe that DMUs consist of divisions and the production process takes place at a division level. In this circumstance, it is beneficial to analyze division-level efficiency. However, since it may often be difficult or impossible to attribute shared inputs to specific divisions, input data are less likely to be available at a division level than are (intermediate and final) output data. In **case 1** (or **case 3** in Appendix 1), we can modify the second-stage equation into J equations (where J is the number of divisions of each DMU) and then estimate second-stage efficiency of each division (see Appendix 2 for model specification and efficiency calculation, and Section 7 for an empirical example).

On the other hand, we might face another scenario where data on input and intermediate outputs are disaggregated and data on final outputs are aggregated. For example, in infrastructure maintenance, we often have data on labor and materials (inputs) and activity volumes (intermediate output) by maintenance types (such as renewals, repair, and replacement); but no breakdown of asset quality (final output) by maintenance types as it is a result of all maintenance activities carried out with respect to the asset. Given this, in **case 2**, we can transform the first-stage equation of the system of Eqs. (5) – (6) into J equations, each equation models the first stage production process of each division. Doing so yields insight into the performance of divisions at the first stage (See Appendix 2 for detail).

5 Multi-step estimation procedure

Although Network SFA offers several advantages, it faces computational challenges since it is a system of stochastic frontier equations, for which full-information maximum likelihood approaches to estimation may become very complex. This may be one of reasons for its lack of development in the literature to date. We propose a multi-step estimation method which is straightforward to implement and performs comparably to that of Huang et al (2017, 2018). Our estimator can be considered an extension of the corrected ordinary least squares (COLS) estimator which was developed by Olson et al. (1980) for conventional single-stage SFA. A comparison of the performance between COLS and MLE for single-stage SFA can be found in Olson et al. (1980), Coelli (1995), Andor and Hesse (2014), and Andor and Parmeter (2017).

Our estimation includes three steps.

The first step is to estimate slope coefficients. If the system of equations suffers from no endogeneity[17], the SUR estimator is consistent regardless of the correlation between the composite errors across the equations. Thus, it can be used to gauge the slope coefficients. In the presence of the dependency between the errors, SUR estimation increases the precision of the estimator by simultaneously estimating parameters of the system while taking account of the correlation.

In the presence of endogeneity, e.g. Equations (1)-(2) (**case 1**) and (5)-(6) (**case 2**), SUR estimates are no longer consistent. We will need to use an IV estimator. Before explaining the estimation method, it is worth mentioning that our systems meet the order condition of identification. The condition requires that the number of endogenous variables (on the right hand side) in an equation is not greater than a total of the number of exogenous variables in the other equations. The endogenous variables (appearing on the right hand side of the system) are intermediate outputs in both cases. For **case 1**, the exogenous variables are final outputs. Since the number of final outputs in Eq. (2) is at least as great as the number of intermediate outputs in Eq. (1), the condition holds. For **case 2**, the exogenous variables are inputs. Since the number of inputs in Eq. (5) is not less than that of intermediate outputs in Eq. (6). The system of (5)-(6) satisfies the condition.

The IV estimator - 3SLS - is used to estimate the slope coefficients.[18] 3SLS (Zellner and Theil, 1962) is a natural choice here, since we have good reason to believe that the errors are correlated across equations for a given firm. Given this, 2SLS is inefficient, and control function approaches (Heckman and Robb 1985, Wooldridge 2015) would be inappropriate, since they require strict exogeneity of the residuals from the first-stage residuals. Full information maximum likelihood (FIML) (Amemiya 1977) is another alternative that could offer further efficiency gains if the model is correctly specified, but at the cost of considerably additional complexity in our setting, where error distribution departs from multivariate normality.

3SLS can be understood as a combination of two-stage least squares and generalized least squares. This enables 3SLS to address the endogeneity in the system while accounting for the correlation of the error terms across equations. Note that the intercepts generated in the first step need to be corrected since it is based on the assumption that errors have zero means.

The second step is to correct the intercepts and estimate distributional parameters. For simplicity's sake, we will use the method of moments (instead of copula-based MLE) in

[17] such as the system of **case 3** in Appendix 1.

[18] See Hausman, 1983 and Greene, 2012 for estimation method of simultaneous equations.

this step. Following Huang et al. (2017, 2018) we assume that the dependence between the composite errors can be captured by a copula. Therefore, it is possible to identify the marginal distribution of a composite error in an equation without specifying its dependence with composite errors in other equations, and thus the moments of composite errors can be easily derived. In addition, we assume that (i) inefficiency follows a half-normal distribution, $u_{ki} \sim N^{+}(0, \sigma_{uk}^2)$, $k = 1, \ldots, K$ where k denotes a specific equation of the system of K equations and (ii) noise follows a normal distribution, $v_{ki} \sim N(0, \sigma_{vk}^2)$. Under these assumptions, σ_{uk}^2 and σ_{vk}^2 can be determined using the following equations (Olson et al. 1980)

$$\frac{1}{n-1}\sum \left(\hat{\varepsilon}_{ki}^{*}\right)^3 = \hat{m}_{3k} = (-1)^h \sqrt{\frac{2}{\pi}}\left(\frac{4}{\pi} - 1\right)\hat{\sigma}_{uk}^3 \tag{9}$$

$$\frac{1}{n-1}\sum \left(\hat{\varepsilon}_{ki}^{*}\right)^2 = \hat{m}_{2k} = \left(1 - \frac{2}{\pi}\right)\hat{\sigma}_{uk}^2 + \hat{\sigma}_{vk}^2 \tag{10}$$

Where n is the number of observations; $\hat{\varepsilon}_{ki}^{*}$ is the estimated residuals obtained from the first step; $\hat{m}_2$ and $\hat{m}_3$ are the second and third moments of the residuals $\hat{\varepsilon}_{ki}^{*}$; $h = \begin{cases} 0 \;\; if \;\; \varepsilon_{ki} = v_{ki} + u_{ki} \\ 1 \;\; if \;\; \varepsilon_{ki} = v_{ki} - u_{ki} \end{cases}$.

The mean of inefficiency of equation k is

$$\hat{E}\left(u_{ki}\right) = \sqrt{2/\pi}\hat{\sigma}_{uk} \tag{11}$$

Thus, the corrected intercepts and residuals of the system are

$$\hat{\alpha}_k = \hat{\alpha_k^{*}} - (-1)^h \sqrt{2/\pi}\hat{\sigma}_{uk} \tag{12}$$

$$\hat{\varepsilon}_{ki} = \hat{\varepsilon}_{ki}^{*} + (-1)^h \sqrt{2/\pi}\hat{\sigma}_{uk} \tag{13}$$

Where $\hat{\alpha_k^{*}}$ is the estimated intercepts of equation k obtained from the first step.

The last step is to calculate (in)efficiency. Ideally we should use all information to estimate stage inefficiency, i.e $\hat{u}_{ki} = E\left(u_{ki}, |, \hat{\varepsilon}_{1i}, \ldots, \hat{\varepsilon}_{Ki}\right)$ or $\hat{u}_{ki} = M\left(u_{ki}, |, \hat{\varepsilon}_{1i}, \ldots, \hat{\varepsilon}_{Ki}\right)$. But deriving this conditional mean or mode is complex and the aim of the paper is to provide a relatively simple estimation method. Thus, following Lai and Huang (2013) and Huang et al. (2017, 2018) we calculate stage inefficiency $\hat{u}_{ki}$ conditional on $\hat{\varepsilon}_{ki}$ only, using the formula proposed by Jondrow et al. (1982). Finally, overall inefficiency is aggregated from stage inefficiency.

The estimation steps for the systems of Eqs. (1) – (2) and (5) – (6) are showed in Appendix 4.[19]

6 Monte Carlo simulations

This section is to examine the performance of our estimation method in the second step, i.e. the method of moments in comparison with that of the copula-based MLE used in Huang et al. (2017, 2018). It is noted that as the first and third steps of our estimation method are similar to those of Huang et al. (2017, 2018), we will not discuss them in this section.

We will use **case 1** with a simple system of two equations as an illustration. The data generating process is as follows:

$$\text{The first stage equation} : \ln(x) = \alpha_1 + \beta_1 \ln(z) + \varepsilon_1 \tag{14}$$

$$\text{The second stage equation} : \ln(z) = \alpha_2 + \beta_2 \ln(y) + \varepsilon_2 \tag{15}$$

Where $\varepsilon_k = u_k + v_k, k = 1, 2, \quad u_k \sim N^{+}(0, \sigma_{uk}^2)$ and $v_k \sim N(0, \sigma_{vk}^2)$.

The probability density function (PDF) of the composite errors is

$$f_k\left(\varepsilon_k\right) = \frac{2}{\sigma_k}\phi\left(\frac{\varepsilon_k}{\sigma_k}\right)\Phi\left(\frac{\varepsilon_k \lambda_k}{\sigma_k}\right) \tag{16}$$

Where $\phi(.)$ is the standard normal PDF, $\Phi(.)$ is the standard normal cumulative distribution function (CDF), $\sigma_k = \sqrt{\sigma_{uk}^2 + \sigma_{vk}^2}$ and $\lambda_k = \frac{\sigma_{uk}}{\sigma_{vk}}$.

The CDF of the composite errors is

$$F_k\left(\varepsilon_k\right) = \int_{-\infty}^{\varepsilon_k} \frac{2}{\sigma_k}\phi\left(\frac{x}{\sigma_k}\right)\Phi\left(\frac{x\lambda_k}{\sigma_k}\right)dx \tag{17}$$

Following Lai and Huang (2013) and Huang et al. (2017, 2018), we assume that the dependence between the composite errors can be captured by Gaussian copula function C. The joint CDF of the composite errors can be written as

$$\begin{aligned} F\left(\varepsilon_1, \varepsilon_2\right) &= C\left(F_1\left(\varepsilon_1\right), F_2\left(\varepsilon_2\right); p\right) \\ &= \Phi_2\left(\Phi^{-1}\left(F_1\left(\varepsilon_1\right)\right), \Phi^{-1}\left(F_2\left(\varepsilon_2\right)\right); p\right) \end{aligned} \tag{18}$$

Where $\Phi_2(.; p)$ is the CDF of a standard bivatiate normal distribution with the correlation coefficient of p and $\Phi^{-1}(.)$ is the inverse CDF of a standard normal distribution.

Equation (18) shows that $\Phi^{-1}\left(F_1\left(\varepsilon_1\right)\right)$ and $\Phi^{-1}\left(F_2\left(\varepsilon_2\right)\right)$ follow a standard bivatiate normal distribution. Thus, we

[19] We have written a Stata package "networkSFA" to enable estimation of our Network SFA with the multi-step estimation method. See Appendix 7 for detail.

can generate ε_1, ε_2 and the other variables, using the following steps:

Step 1: Draw ζ_1 and ζ_2 from a standard bivatiate normal distribution with a correlation coefficient p.

Step 2: Compute ε_1 and ε_2 as $\varepsilon_1 = F_1^{-1}(\Phi(\zeta_1))$ and $\varepsilon_2 = F_2^{-1}(\Phi(\zeta_2))$.

Step 3: Draw $\ln(y)$ from a standard normal distribution, i.e. $\ln(y) \sim N(0,1)$ and compute $\ln(z)$ and $\ln(x)$ as $\ln(z) = \ln(y) + \varepsilon_2$ and $\ln(x) = \ln(z) + \varepsilon_1$.

The slope parameters β_1 and β_2 are estimated using 3SLS. Then the intercepts (α_1, α_2) and the distributional parameters $(\sigma_{u1}, \sigma_{u2}, \sigma_{v1}, \sigma_{v2})$ are estimated using two different methods. One is the methods of moment we proposed in Section 5. The other is the copula-based MLE used in Huang et al. (2017, 2018).

For the copula-based MLE, we will need to derive the joint PDF of the composite errors. Since the joint PDF is the derivative of the joint CDF expressed in Eq. (18) with respect to $\varepsilon_1, \varepsilon_2$, the joint PDF takes the form

$$f(\varepsilon_1, \varepsilon_2) = \frac{1}{|P|^{1/2}} \exp\left(-\frac{1}{2}\zeta^{'}(P^{-1} - I)\zeta\right) f_1(\varepsilon_1) f_2(\varepsilon_2) \qquad (19)$$

Where P is the correlation matrix $P = \begin{pmatrix} 1 & p \\ p & 1 \end{pmatrix}$, I is a 2×2 identity matrix $I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ and $\zeta^{'} = \left(\Phi^{-1}(F_1(\varepsilon_1)), \Phi^{-1}(F_2(\varepsilon_2))\right)$.

Thus, the log-likelihood function for a sample of N observations is[20]

$$\begin{aligned} \ln L \quad = \quad & -\tfrac{N}{2}\ln|P| - \tfrac{1}{2}\textstyle\sum_1^N \zeta^{'}\left(P^{-1} - I\right)\zeta + \sum_1^N \ln f_1(\varepsilon_1) \\ & + \textstyle\sum_1^N \ln f_2(\varepsilon_2) \end{aligned} \qquad (20)$$

We set $\alpha_1 = \alpha_2 = 0$, $\beta_1 = \beta_2 = 1$, $\sigma_{u1} = \sigma_{u2} = \sigma_{v1} = \sigma_{v2} = 1$. The corresponding distributional parameters of the composite errors are $\sigma_1 = \sigma_2 = \sqrt{2}$ and $\lambda_1 = \lambda_2 = 1$. We set the number of replications of 1,000.[21] We then vary the sample size n and degrees of the dependence p. We consider cases $n = 250$, $n = 500$, $n = 1000$ and $p = -0.5$, $p = 0$, $p = 0.5$.

The estimates of the intercepts and the distributional parameters by the two estimation methods are summarized in Table 2. The Monte Carlo simulation shows that the estimates produced by the two methods aare very similar. Our estimator shows superior performance in small sample sizes while the copula-based MLE performs better with large sample sizes. Thus, our proposed method, the method of moment is not inferior to the copula-based MLE but offers a significantly simpler implementation.

7 Empirical example of highway maintenance costs in England

In this section, we will apply the model developed in Section 4 to assess performance of English Local Highway Authorities (LHAs) who are responsible for maintaining non-trunk A roads and all B, C, and U roads in their areas.[22] Since the LHAs are highly regulated, we will start by discussing the optimal behaviors of regulated firms before the application.

7.1 Optimal behaviors of regulated firms

Given our focus is to evaluate the performance of network industries, we now reflect on the behaviors of these firms whose performance is of great interest to regulatory authorities.

We consider that input-orientation is most appropriate and we illustrate this with two examples.

Example 1: For utility firms such as electricity distribution network operators and sewage and water companies, these firms have responsibility for supplying essential goods like electricity, water, and gas to the public. Unlike conventional firms, their operation relies on the transmission infrastructures which are usually built by the government due to their huge initial costs. Because of their important role in people's daily lives, and their strong market power, utility firms are highly regulated by regulation authorities. They are restricted on what they can and cannot do. For example, British water and electricity companies are not allowed to deny joining of customers in their supply region. They also cannot provide products and services outside their prescribed region. Therefore, their final outputs tend to be fixed and beyond their direct control. For this reason, utility firms are likely to have fixed final outputs. Under the assumption

[20] The parameters of Eq. (20) are estimated using the R package *maxLik*. We use true parameter values as starting values to avoid overstate performance of our MM estimator in comparison with MLE due to suboptimal starting values. On average, the optimization of the log-likelihood function takes approximately 20 s.

[21] We start with 1,500 replications and exclude those with wrong skewness. The issue of wrong skewness reduces as the sample size increases (i.e. approximately 35, 20, and 10% of the replications have wrong skewness for the sample size of 250, 500, and 1000, respectively). We only keep the first 1000 correct-skewness replications. This is because our method-of-moments estimator is not applicable to wrong skewness (See Olson et al. (1980) for the discussion of the problems of the method-of-moments estimator). Although the copula-based MLE can be used to estimate the distributional parameters, our simulation results suggest that the copula-based MLE seems to be sensitive to wrong skewness. In such cases, the MLE estimates can be considerably different from the true values (e.g. the means of the estimates of σ_{u1} and σ_{u2} are around 0.5 for wrong skewness replications in our simulations).

[22] The definition of A, B, C and U roads is provided by the UK government (the Department of Transport, 2012).

Table 2 The estimates of the intercepts and the distributional parameters

Sample size	Parameters	True value	$p = 0.5$					$p = 0$					$p = -0.5$				
			Mean		RMSE			Mean		RMSE			Mean		RMSE		
			MM	CB MLE	MM	CB MLE	Ratio	MM	CB MLE	MM	CB MLE	Ratio	MM	CB MLE	MM	CB MLE	Ratio
***n*=250**	σ_{u1}	1	1.048	1.025	0.292	0.322	0.906	1.043	1.045	0.292	0.301	0.970	1.021	1.027	0.291	0.297	0.978
	σ_{u2}	1	1.037	1.013	0.280	0.323	0.869	1.049	1.050	0.286	0.301	0.951	1.034	1.034	0.282	0.304	0.929
	σ_{v1}	1	0.961	0.963	0.127	0.125	1.017	0.970	0.965	0.116	0.119	0.976	0.974	0.969	0.123	0.121	1.019
	σ_{v2}	1	0.965	0.966	0.113	0.117	0.968	0.960	0.955	0.122	0.127	0.962	0.969	0.963	0.110	0.113	0.973
	α_1	0	−0.037	−0.019	0.244	0.266	0.918	−0.030	−0.032	0.254	0.261	0.976	−0.019	−0.024	0.256	0.261	0.983
	α_2	0	−0.026	−0.008	0.234	0.269	0.872	−0.042	−0.043	0.236	0.248	0.951	−0.028	−0.028	0.236	0.253	0.936
***n*=500**	σ_{u1}	1	1.008	0.997	0.241	0.254	0.946	1.001	1.003	0.247	0.250	0.986	1.002	1.000	0.240	0.255	0.943
	σ_{u2}	1	1.013	1.004	0.236	0.246	0.956	1.001	0.999	0.248	0.256	0.969	1.004	1.001	0.242	0.254	0.956
	σ_{v1}	1	0.987	0.988	0.093	0.090	1.024	0.987	0.985	0.086	0.086	0.997	0.988	0.986	0.090	0.087	1.037
	σ_{v2}	1	0.982	0.983	0.087	0.086	1.014	0.985	0.984	0.089	0.090	0.987	0.985	0.984	0.088	0.088	1.004
	α_1	0	−0.002	0.007	0.201	0.212	0.947	0.002	0.000	0.208	0.211	0.988	−0.001	0.001	0.204	0.216	0.945
	α_2	0	−0.009	−0.002	0.195	0.203	0.956	0.000	0.001	0.203	0.209	0.971	−0.003	−0.002	0.200	0.207	0.963
***n*=1000**	σ_{u1}	1	0.979	0.980	0.204	0.200	1.020	0.967	0.970	0.205	0.203	1.008	0.967	0.972	0.207	0.198	1.044
	σ_{u2}	1	0.975	0.978	0.201	0.190	1.056	0.976	0.976	0.214	0.211	1.011	0.976	0.981	0.210	0.198	1.064
	σ_{v1}	1	0.996	0.995	0.072	0.069	1.043	1.003	1.001	0.068	0.069	0.998	1.003	1.001	0.072	0.068	1.055
	σ_{v2}	1	0.999	0.998	0.065	0.062	1.050	0.998	0.997	0.071	0.070	1.006	0.997	0.996	0.070	0.066	1.050
	α_1	0	0.015	0.015	0.168	0.164	1.021	0.028	0.025	0.172	0.170	1.007	0.027	0.023	0.175	0.169	1.033
	α_2	0	0.020	0.018	0.162	0.153	1.059	0.019	0.019	0.172	0.170	1.011	0.019	0.015	0.171	0.161	1.064

"MM" is the method of moment, "CB MLE" is copula-based MLE, RMSE is the root mean squared errors, and Ratio is the ratio of RMSEs of MM to RMSEs of CB MLE. It is noticed that $\hat{\beta}_1$ and $\hat{\beta}_2$ of the two methods are the same, thus their results are not shown here of rational behavior, they will try to minimize intermediate outputs and inputs, given their fixed final outputs. Therefore, the behavior assumption of input-orientation (and thus the corresponding model specification) is suitable for utility firms.

Example 2: Similarly, organizations whose duties are to maintain quality of public assets such as roads and railway systems or community assets seem to have fixed final output. They have to ensure that the assets meet safety and quality standards for use. Thus, they seem to have target asset quality (i.e. final output) that they aim to maintain.[23] Given the target, they optimize their intermediate outputs and inputs. This matches the behaviors described in case 1. The SFA with input-orientation should be used to evaluate the performance of these organizations.

7.2 Model and data of road maintenance

As discussed in the introduction, the maintenance process of LHAs can be split into two stages. In the first stage, LHAs incur costs to implement road treatments such as overlaying, surfacing and dressing which then help to improve road condition in the second stage (Fig. 2). A similar framework is used by Rouse and Chiu (2009) to describe the production process of road maintenance. The authors refer it as an input–activity–output process. It is worth mentioning that these two stages occur simultaneously rather than sequentially. This simultaneity enhances interdependence between the stages. For example, the choice of treatment type not only impacts procurement and resource allocation in the first stage but also influences the effectiveness of the treatments in achieving the desired road condition in the second stage.

Despite a few works investigating efficiency of road maintenance, their focus was not devoted to this multi-stage feature. For example, Fallah-Fini et al. (2012), Ozbek et al. (2012) and Kalb (2014) used DEA to analyze the performance of road maintenance operations in US and Germany. Wheat (2017), Wheat et al. (2019), Stead and Wheat (2020) and Yarmukhamedov et al. (2020) on the other hand employed SFA to predict efficiency in road maintenance in England and Sweden. All of these studies used single-stage models to measure (overall) efficiency. To the best of the authors' knowledge, only Rouse and Chiu (2009) incorporated a two-stage production process in their performance evaluation of road maintenance in New Zealand, using DEA.

We thus contribute to the limited literature on road maintenance by using Network SFA to measure overall and stage efficiencies of LHAs. Along with the advantage of explicitly considering the two stages, the model meets the need of exploiting different aggregations of LHAs' data where disaggregated data on outputs but aggregated data on costs are available.

To perform the analysis, we use an unbalanced panel dataset of 142 LHAs from the financial year 2009/10 to 2017/18. This provides us with 755 LHA-year observations in total. Our cost data is aggregated at the LHA level but the (intermediate and final) output data can be disaggregated by road types. In additional to data availability, it appears that LHAs have different standards, expectations and behaviors on how they maintain A roads versus other roads, implying different frontiers for different road types. Thus, we separate LHAs' road networks into two road types: major roads including non-trunk A roads and minor roads including all B, C, and U roads. We will evaluate the performance of LHAs by road types where possible.

We define total costs as a sum of annual revenue expenditures and capital expenditures associated with road maintenance activities (excluding new road construction costs). For input prices, we do not use LHA-level data. As an alternative, we can use median hourly wages of the civil engineering sector by regions and a national index of material prices for road construction as proxies for labor and material prices. However, the latter is invariant across LHAs and will be (partially) captured by our time trend variable. Thus, we decide to deflate the costs by median hourly wages and consider our first-stage equation an input requirement function where the input is an aggregate input expressed in the monetary value (i.e. cost) of all factors used in the production process. Regarding outputs, we use road kilometers receiving treatments as a measure of the intermediate outputs. Meanwhile, the final output is defined as changes in road condition from previous year, where road condition is a proportion of road without a need of maintenance.[24]

For control variables, following the literature (Stead and Wheat 2020 and Yarmukhamedov et al. 2020) we include traffic density (in both stages) as it causes wear-and-tear damage, requiring greater treatment (in the second stage)

Fig. 2 Production process of English LHAs

[23] We note our discussion under Case 3 in Appendix 1 where we state for some specific assets a quality metric is not easy to monitor and so that case might be more appropriate for those specific assets. However, at a more aggregate company level, input-orientation is more appropriate.

[24] We could use road condition as a final output and control previous-year road condition. However, due to their high correlation (the correlation coefficient is 0.9), we decided to use their difference to avoid multi-collinearity.

Table 3 Descriptive statistics of the variables

Variable	Unit	Mean	Std. dev.	Min	Max
$cost$	£ thousand	1,339.46	1,232.09	109.10	12,876.80
RT^M	Km	18.34	25.75	0.20	160.20
RT^m	Km	98.98	161.07	0.73	1,318.00
RL^M	Km	251.18	266.03	12.85	1,052.40
RL^m	Km	2,145.53	2,359.19	44.20	11,866.40
RC^M	Ratio	0.95	0.03	0.68	1.00
RC^m	Ratio	0.86	0.08	0.33	0.99
TD^M	Thousand vehicles/km	5,036.83	1,719.05	1,663.26	10,922.62
TD^m	Thousand vehicles/km	701.30	263.02	125.69	1,482.74
$land$	Hectare	109,839.30	164,823.80	290.39	803,771.60

and increases costs related to access and possessions of the network (in the first stage). We also include a time trend to capture technical change.[25] Finally, we control for the size of LHAs (represented by their land area) in the first stage and the network length in the second stage. We expect that the size can either positively or negatively impact maintenance costs. A small size might imply that the network is in an urban area with extensive underground assets, requiring greater efforts and costs of maintenance. However, a network spreading over a small area could incur less costs of material transport. To check if there is a nonlinear relationship between the size and costs we include land area (in a log form) and its square. For the second-stage equations, because intermediate outputs (the dependent variables) are measured as kilometers of road treatment (in a log form) and the final outputs (the explanatory variables) are the change in the proportion of roads in good condition, there is a need to control for network length. Given the same road condition change in the proportion, bigger networks require bigger treatment amounts. (See Appendix 5 for definitions and sources of the variables).

Because LHAs have to ensure that their road network is fit-for-purpose it is reasonable to assume that they have target road conditions. Given this target, LHAs determine maintenance works they need to do and costs incurred. Thus, the final outputs are fixed and input-orientation is the most appropriate. Following the literature on English road maintenance (Wheat 2017, Wheat et al. 2019, Stead and Wheat 2020), we assume that final output (i.e. road condition) is exogenous. Therefore, **case 1** with different data aggregation is used. The equation system (1) – (2) is rewritten for our specific case of road maintenance as

Stage 1:

$$\begin{aligned}\ln\left(cost_{it}\right) = \; & \alpha_1+\beta_1^M\ln\left(RT_{it}^M\right)+\beta_1^m\ln\left(RT_{it}^m\right)\\ & +\gamma_1\ln(TD_{it}^M)+\gamma_2\ln(TD_{it}^m)+\gamma_3\ln\left(land_{it}\right)\\ & +\gamma_4(\ln\left(land_{it}\right))^2+\gamma_5 t+\gamma_5 t^2+u_{1it}+v_{1it}\end{aligned} \tag{21}$$

Stage 2:

$$\begin{aligned}\ln\left(RT_{it}^M\right) = \; & \alpha_2^M+\beta_2^M\triangle RC_{it}^M+\omega_1^M\ln(RL_{it}^M)\\ & +\omega_2^M\ln(TD_{it}^M)+\omega_3^M t+u_{it}^M+v_{it}^M\end{aligned} \tag{22}$$

$$\begin{aligned}\ln\left(RT_{it}^m\right) = \; & \alpha_2^m+\beta_2^m\triangle RC_{it}^m+\omega_1^m\ln(RL_{it}^m)\\ & +\omega_2^m\ln(TD_{it}^m)+\omega_3^m t+u_{it}^m+v_{it}^m\end{aligned} \tag{23}$$

where i and t represent LHAs and time periods $i=1,\ldots 142,\; t=0,\ldots,8$ with $t=0$ corresponding to the financial year 2009/10, M and m denotes major roads and minor roads respectively, $cost$ is total maintenance costs, RT is road treatment kilometers, RC is road condition measured as $\frac{Road\ length\ without\ need\ of\ maintenance}{Total\ road\ length}$, $\triangle RC_{it}=RC_{it}-RC_{i(t-1)}$, RL is road length, TD is traffic density, and $land$ is land area.[26]

Summary statistics of the variables are shown in Table 3. As seen in Table 3, there is a considerable difference between major roads and minor roads where major roads tend to have better road condition despite their high traffic density.

The system of Eqs. (21)–(23) is estimated, using the multi-step procedure in Section 5. Since the system has the endogeneity, 3SLS will be employed to estimate slope coefficients of the system. For stage efficiency prediction, we will use the conditional mode $\hat{u}_{ki}=M\left(u_{ki},|,\hat{\varepsilon}_{ki}\right)$ as it can be understood as a maximum likelihood estimator of inefficiency, given the composite errors. Furthermore, using the conditional mean results in no observations lying on efficient frontiers, i.e. no LHA is fully efficient at any stage. As overall inefficiency is a sum of (scaled) stage inefficiency this then leads to high overall inefficiency, i.e. overall inefficiency is further from zero, compared to stage inefficiency,

[25] The squared term of the time trend is excluded in the second-stage equations because it is insignificant at a 10% significance level (p-value>0.2).

[26] If we use the change in the logarithm of kilometers of roads in good condition as a proxy for the final output, i.e. $\triangle RC_{it}=\ln\left(Road\ length\ without\ need\ of\ maintenance_{it}\right)-\ln(Road\ length\ without\ need\ of\ maintenance_{i(t-1)})$, the results are very similar.

which is not desirable. We also estimate a single-equation SF model for comparison.

7.3 Estimated frontiers for road maintenance

Parameter estimates from the Network SF and single-stage SF models are shown in Table 4.

We first discuss the frontier results. All coefficients which are statistically significant have the expected signs. They all meet the theoretical expectation that the output variables have positive and significant elasticities in the cost and input requirement functions.

For the first production stage, we observe economies of scale in road treatment volume. The elasticity of cost with respect to treatment length is 0.72 (0.43 + 0.29).[27] This suggests that by increasing maintenance work length LHAs are likely to reduce their unit costs. A positive impact is also found for traffic density since higher usage increases road pavement damage, as well as hindering ease of access to the infrastructure, both factors increasing maintenance costs.

We find a non-linear effect of land area (Table 5 and Fig. 3), implying that there is a cost-minimizing land area. This may be because small land areas correspond to urban areas where the prevalence of utilities and other underground assets implies work on the network is expensive. Large land areas on the other hand imply extra costs for material transport for example. Our model implies that the optimal land area is around 123,000 hectares.[28] Regarding time effects, costs increase in the first three years of the examined period and then decrease for the remaining years.[29] This would indicate that the efficient cost of road maintenance has fluctuated over time, and it is not the case that there has been a positive frontier shift at least in the first half of the sample.

Moving onto the second production stage, again we find a positive relationship between final outputs and intermediate outputs. Our estimation results suggest that improving road conditions by one percent requires a proportionally greater increase in road treatments for major roads (estimated at 1.66) compared to minor roads (estimated at 1.04). This finding is reasonable, as major roads typically start in better conditions, meaning that some treatments may be applied to already well-maintained sections to prevent future deterioration. Increased

Table 4 Stochastic frontier estimated parameters

Network SFA	Coeff.	SE	Single-stage SFA	Coeff.	SE
Stage 1: ln(cost)					
$\ln(RT^M)$	0.43***	0.11	$\triangle RC^M_{it}$	1.003	0.91
$\ln(RT^m)$	0.29***	0.10	$\triangle RC^m_{it}$	−0.002	0.50
$\ln(TD^M)$	0.29***	0.11	$\ln(TD^M)$	0.40***	0.10
$\ln(TD^m)$	0.33***	0.09	$\ln(TD^m)$	0.34***	0.08
$\ln(land)$	−1.43***	0.17	$\ln(RL^M)$	0.72***	0.10
$\ln{(land)}^2$	0.06***	0.01	$\ln(RL^m)$	0.26**	0.11
t	0.07**	0.04	$\ln(land)$	−1.42***	0.17
t^2	−0.02***	0.00	$\ln{(land)}^2$	0.06***	0.01
intercept	8.59***	1.04	t	0.02	0.03
Stage 2 – Major road: ln(RT)^M			t^2	−0.01***	0.00
$\triangle RC^M_{it}$	1.66**	0.77	intercept	3.69***	0.92
$\ln(TD^M)$	0.24***	0.08			
$\ln(RL^M)$	1.24***	0.03			
t	0.01	0.01			
intercept	−6.29***	0.75			
Stage 2 – Minor road: ln(RT)^m					
$\triangle RC^m_{it}$	1.04**	0.53			
$\ln(TD^m)$	−0.05	0.06			
$\ln(RL^m)$	1.22***	0.03			
t	0.00	0.01			
intercept	−4.80***	0.54			

*, **, *** denotes significance at the 10%, 5% and 1% level, respectively. Reported intercepts for the Network SF model are $\hat{\alpha}^* = \hat{\alpha} + \hat{E}\,(u)$. Estimated intercepts from the copula-based MLE are presented in Appendix 6. We do not report the slope coefficient estimates from the multi-step copula-based MLE as they are identical to those in Table 4. This is because the first step which estimates slope coefficients is the same across the two approaches. Additionally, we do not report the dependence parameter, as our estimation method does not allow for its estimation

Table 5 Elasticities of cost with respect to land area from the Network SF model

Variable	Mean	Std. dev.	Min	Max
Elasticities _land	−0.17	0.21	−0.74	0.23

traffic density leads to increased maintenance requirements on major roads as expected but no impact is found on minor roads. This suggests that traffic density on minor roads, which is seven times lower than that on major roads on average, has not reached the point where it is a major component of damage to the road pavement vis-à-vis time related factors such as weathering. We also find that road treatment increases with road length and the time trend was estimated as statistically insignificant in the second stage. This implies that there is little evidence that efficient treatment frontiers shift over the examined period.

[27] The 95% confidence interval of the elasticity is (0.57, 0.88). The upper bound is less than 1 indicating that constant returns to scale can be rejected at the 5% significance level.

[28] The 95% confidence interval of the optimal land area is (36,000; 427,000).

[29] The value of t at which costs are maximum is −0.07/(2*0.02) = 2.09. Note that t starts from zero. Thus, costs increase in the first three years before decreasing.

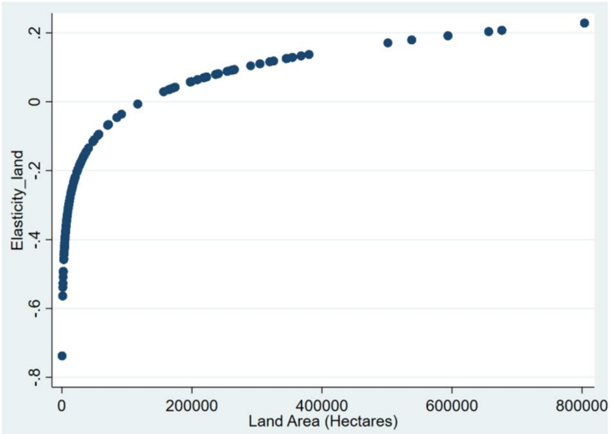


Fig. 3 Scatter diagram of cost elasticities with respect to land area

From our results, we can calculate elasticity of cost with respect to the final outputs.[30] Our model predicts it is 1.02 (=0.43*1.66+0.29*1.04) which is not statistically significantly different from 1.[31] This implies constant returns to scale overall.

Comparing these results to those from single-stage SFA, we see both commonalities and key differences. For the control variables (traffic density, road length, land areas and time trends), their effects on maintenance costs are very similar between the two approaches. However, the biggest difference between the two approaches is that the impact of final outputs on cost is insignificant in the single-stage SFA. Whilst overall, the estimate of returns to scale is similar, now being 1.00 (=1.00+0.00), the difference between the impact of minor and major road conditions is implausible (effectively zero versus one). Put another way, the single-stage SFA suggests that maintenance costs are hardly impacted by changes in the final output of minor roads (the coefficient of −0.002) but increase proportionately with an increase in the final output of major roads (the coefficient of 1.003). Whilst it may be true that condition changes on minor roads might impact cost less than changes on major roads, the magnitude of the difference raises concerns about its plausibility. We attribute this difference to Network SFA exploiting the disaggregated intermediate and final output data enabling two Stage 2 equations to be estimated and the impact of major road treatments to be considered separately to those on minor roads. We thus flag the importance of our approach in exploiting disaggregate data available at some but not all stages as a major benefit of our approach.

[30] The estimated parameters associated with the final outputs can be interpreted as elasticities because the final outputs are expressed in a ratio form.

[31] Where 1.66 and 1.04 are the estimated coefficients of $\triangle RC_{it}^{M}$ and $\triangle RC_{it}^{m}$ in the second and third equations (as shown in Table 4).

Table 6 Descriptive statistics of estimated efficiency

Methods	Efficiency	Mean	Std. dev.	Min	Max
Method of moments	First-stage efficiency	0.68	0.24	0.11	1.00
	Second-stage efficiency (Major)	0.83	0.16	0.29	1.00
	Second-stage efficiency (Minor)	0.89	0.11	0.35	1.00
	Overall efficiency	0.59	0.19	0.11	1.00
Copula-based MLE	First-stage efficiency	0.69	0.23	0.12	1.00
	Second-stage efficiency (Major)	0.91	0.10	0.51	1.00
	Second-stage efficiency (Minor)	0.96	0.04	0.69	1.00
	Overall efficiency	0.65	0.21	0.12	1.00
Single-stage	Overall efficiency	0.66	0.23	0.12	1.00

The table presents summary statistics of estimated efficiency i.e. $\exp[-M\,(u_{ki}, |, \hat{\varepsilon}_{ki})]$ for stage efficiency and $\exp\left(-\left(\hat{u}_{1it} + \hat{\beta}_1^M \hat{u}_{it}^M + \hat{\beta}_1^m \hat{u}_{it}^m\right)\right)$ for overall efficiency

7.4 Predicted efficiency

After the frontier and distributional parameters are estimated, we calculate stage inefficiency using the conditional mode. Overall efficiency is then calculated from the stage efficiencies as

$$\exp\left(-\hat{u}_{it}\right) = \exp\left(-\left(\hat{u}_{1it} + \hat{\beta}_1^M \hat{u}_{it}^M + \hat{\beta}_1^m \hat{u}_{it}^m\right)\right).$$

The predicted efficiencies and their correlations are presented in Tables 6, 7. Comparing efficiency across the stages, the Network SF results suggest that LHAs are less efficient in the first stage. Thus, they might benefit from paying more attention to cost optimization, given required road treatment. Furthermore, we find a negative correlation between stage efficiencies. The correlation coefficients are −0.46 (for major road) and −0.31 (for minor road) which are significant at the 1% level. It means that LHAs who are efficient in the first stage are less likely to be efficient in the second stage. This might be because treatments that are optimal for road conditions appear to be expensive. Besides treatment types, the choice of treatment locations could be another explanation. If LHAs decide to maintain areas that are in relatively good condition, it is unlikely that they will incur high costs nor considerably improve road condition.

For the correlation of efficiencies with respect to the two categories of roads, we find positive dependence with a coefficient of 0.36 (significant at a 1% significance level). This implies that LHAs that are inefficient in maintaining major roads tend to be inefficient in maintaining minor roads. Thus, if there is scope for LHAs to improve their performance on major roads, there is likely room for their improvement on minor roads, although the value of 0.36 indicates this correlation is relatively weak.

Table 7 Correlation matrix between efficiency

Methods	Efficiency	1.	2.	3.	4.	5.	6.	7.	8.
Method of moments	1.First-stage	1.00							
	2.Second-stage (Major)	−0.46	1.00						
	3.Second-stage (Minor)	−0.31	0.36	1.00					
	4.Overall efficiency	0.94	−0.17	−0.10	1.00				
Copula- based MLE	5.First-stage	1.00	−0.46	−0.30	0.94	1.00			
	6.Second-stage (Major)	−0.44	0.99	0.37	−0.14	−0.44	1.00		
	7.Second-stage (Minor)	−0.29	0.36	0.99	−0.08	−0.29	0.37	1.00	
	8.Overall efficiency	0.98	−0.32	−0.22	0.98	0.98	−0.29	−0.20	1.00
Single-stage	9.Overall efficiency	0.78	0.05	0.10	0.90	0.78	0.06	0.10	0.85

Overall efficiency scores obtained from our Network model are comparable to those from the single-stage model, with a high correlation coefficient of 0.9 between them. However, our Network SF model predicts that there are two year-LHA[32]observations efficient at both stages while this figure from the single-stage model is considerably greater – 91 observations.[33] Thus, overall efficiencies from the single-stage SFA seem to be overestimated as a result of the failure of incorporating the multi-stage production process (and thus intermediate outputs) into the model.

To reinforce that our simplified estimation in step two (method of moments) does not distort predictions, we have applied copula-based MLE of Huang et al. (2017, 2018) and re-estimate the intercepts, distributional parameters and efficiency of the Network SF model. The results are very similar.[34] The smallest correlation coefficient between efficiency predicted by the two methods is 0.98 (Table 7) (See Appendix 6 for the result of the estimated intercepts and distributional parameters).

Using the example of road maintenance, we illustrate the benefits of Network SFA. It provides more informative indicators of performance. By employing the Network SF model, we know how LHAs perform not only in the whole production process but also in each stage and on each road type. Furthermore, as it explicitly models the production process of each stage, the Network SF model provides insight into how inputs, intermediate outputs and final outputs are transformed (i.e. economies/returns of scale in each stage). In our empirical example, our simplified multi-step estimation method produces very similar results to those obtained from the copula-based MLE. This result is consistent with the findings from the Monte Carlo simulations presented in Section 6.

8 Conclusion

This study develops a Network SF model which analyzes the production of intermediate outputs and their transformation into final outputs as a system of structural equations.

This approach allows for the prediction of stage-specific efficiency scores and the derivation of the overall technical efficiency in multi-stage production processes, and exploits data on intermediate outputs, offering greater insight into the production process. The key benefit of our Network SF model is to provide greater understanding as to which stage of the production process there exists the greatest opportunity for a DMU to make efficiency improvements. As such our approach can be thought of as helping with the vexing question of why a DMU might be found to be overall inefficient (beyond simply measuring this), at least in the sense of identifying which part of the production process is mainly responsible for inefficiency.

A further advantage is that our model enables different data aggregations to be used in different equations. For example, if data on intermediate outputs and final outputs are available at division level, the second stage equation can be expanded to J equations. Each equation models the second-stage production process of each division. It thus yields performance indicators of not only DMUs but also divisions. In short, the Network SF model exploits the availability of intermediate output data to provide more details about the production process and performance of DMUs.

We propose a straightforward multi-step estimation approach. In the first step, the slope parameters are estimated using SUR (if the system has no endogeneity) or 3SLS (otherwise). The second step estimates intercepts and distributional parameters using a method of moments approach. The final step predicts stage-specific and overall efficiencies. We include an empirical application and Monte

[32] The term "two year LHA observations" means two observations in the panel dataset. These could be the same LHA over two time periods or two different LHAs.

[33] The corresponding figure from the copula-based MLE method is 4 year-LHA observations.

[34] Although the average second-stage efficiency estimates from the method of moments is lower than those from the copula-based MLE in our empirical example (as shown in Table 6), it is not always observed in replications of the Monte Carlo simulations (discussed in Section 6). This, combining with the high correlation in estimated efficiency between the two estimation methods (presented in Table 7), alleviates the concern that the method of moments systematically underestimates efficiency of the second stage.

Carlo simulations, and this simpler approach produces similar results to the copula-based method proposed by Huang et al. (2017, 2018). A Stata package is provided for ease of implementation.

Our empirical application is to English local highways maintenance data. We find evidence of increasing returns to scale in the first stage but constant returns to scale in the overall production process. Our results suggest that overall technical inefficiency in road maintenance is accounted for mainly in the production of intermediate outputs – in our application, volumes of road treatments – rather than in the translation of road treatments into improved road conditions. Thus, LHAs should pay more attention to cost optimization in the first stage. Furthermore, our findings suggest a positive correlation between efficiencies with respect to the maintenance of major and minor roads, but a negative correlation between first-stage and second-stage efficiencies. In comparison with the single-stage model, the Network SF model produces very similar estimated overall efficiency (with the correlation coefficient of 0.90). However, the Network SF model provides a more plausible relationship between cost, intermediate output and final output than the single-stage model and we attribute that to our model being able to better exploit disaggregated intermediate and final output data. Finally, to demonstrate the validity of our proposed estimation method, we re-estimate efficiency using the copula-based MLE in the second step. The results from the two methods are very comparable (with the correlation coefficient of 0.98).

Our modeling framework can be extended to accommodate more complex network structures (without a change in the estimation method) such as a J multi-stage production process (Appendix 3), additional independent inputs entering the second stage for case 2 (and case 3), outputs leaving the production process before the final stage for case 1 (and case 3), multi final outputs for case 1, multi inputs for case 2 (and multi intermediate outputs for case 3). However, there remains much room for further extensions, e.g. to accommodate shared inputs, dynamic specifications with time lags between the production of intermediate and final output. Development of one-step estimation methods is another fruitful avenue for further research. Finally, whilst our application is to panel data, we use a pooled cross-sectional approach to estimation. Extending Network SFA to better exploit panel data would be beneficial.

Appendix

Appendix 1: Model specification under case 3

Case 3: The intermediate output is exogenous (the input and final output are endogenous). In addition, we assume that input-oriented and output-oriented efficiency is pursued in the first and second stage respectively. This implies DMUs have target intermediate output. They then try to minimize input in the first stage and maximize final output in the second stage. Thus, the two-stage SFA model can be specified as

Table 8 The endogeneity in case 3

Behavioral assumption	Two-stage SFA	Single-stage input oriented SFA	Single-stage output oriented SFA
Case 3: Intermediate output is exogenous	No	Yes	Yes

"Yes" means the endogeneity present in the model. "No" means there is no endogeneity in the model

Stage 1:

$$\ln(I_i) = \alpha_1 + \beta_1 \ln(IO_i) + \sum_1^M \gamma_m Z_{1mi} + u_{1i} + v_{1i} \tag{24}$$

Stage 2:

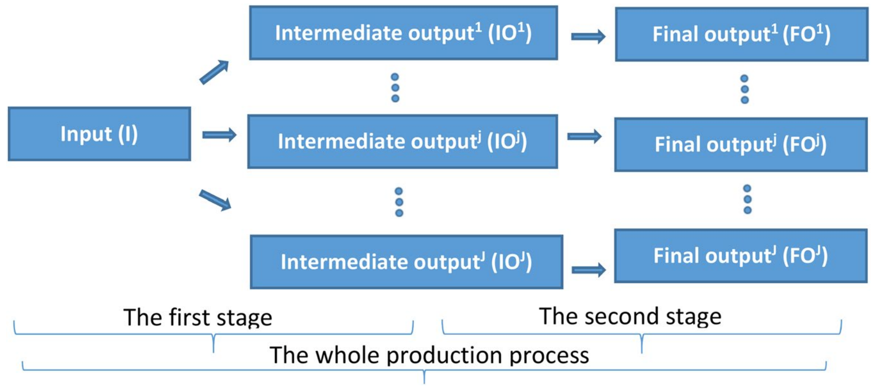
$$\ln(FO_i) = \alpha_2 + \beta_2 \ln(IO_i) + \sum_1^N \omega_n Z_{2ni} - u_{2i} + v_{2i} \tag{25}$$

We can derive input-oriented overall inefficiency by replacing $\ln(IO_i)$ in Eq. (24) with a transformation of Eq. (25). However, this overall (in)efficiency contradicts our orientation assumption since it implies that DMUs have fixed final output. Similarly, we can derive output-oriented overall inefficiency but it will have the same issue. Therefore, a new measure of overall efficiency is needed for fixed-intermediate-output orientation.[35] However, we will not develop that measure in this paper and leave it for future studies. A simple solution for case 3 to avoid the need for new overall efficiency is to use the orientation assumption of either case 1 or case 2.

From a policy perspective, fixed-intermediate-output orientation might appear odd, as society fundamentally derives value from the final output, so it would be sensible for policy makers and regulators to target final outputs rather than intermediate outputs (i.e. input-orientation such as in Case 1). However for certain complex and heterogenous assets e.g. structures, regulators historically have not been able to measure (at sufficient granular level) final outputs and so are forced to monitor and set targets for intermediate outputs. An example is the monitoring of National Highways in Great Britain where the regulator monitors the volume of certain structure related activities rather than the overall condition of the asset (ORR, 2023, Appendix C). The endogeneity of the model in case 3 is shown in the table below (Table 8).

[35] More generally, a new measure is needed for a system of stochastic frontier equations with mixed orientations where some equations are input-oriented while others are output-oriented.

Fig. 4 The two-stage production of a multi-division DMU with aggregated input



Appendix 2: Exploitation of different data aggregations

- *Scenario 1*: data of input is only available at a firm level, but the data of (intermediate and final) outputs is available at a division level; and firms' behavior is **case 1** or **case 3**.

Suppose there are N DMUs, each of them consists of J divisions. The production process of the DMUs can be broken down into two stages. In the first stage, DMUs use inputs to generate J intermediate outputs. Each intermediate output is then used by each division to produce a final output in the second stage (Fig. 4). Since the shared inputs is not allocable to divisions, the data of input is only available at a DMU level. The data of intermediate output and final outputs is however available at a division level. Thus, performance of the first production stage is evaluated at a DMU level. But the second stage should be evaluated at a division level if it is possible.

Case 1: Final outputs are exogenous (input and intermediate outputs are endogenous)

The Network SF model is a system of $(J+1)$ equations. The first equation models the first stage of production (at a DMU level). It can be specified as

$$\ln(x_i) = \alpha_1 + \sum_{j=1}^{J} \beta_1^j \ln(z_i^j) + u_{1i} + v_{1i} \tag{26}$$

Where $i = 1, \ldots, N$, and $j = 1, \ldots, J$ denote DMUs and divisions respectively, u_{1i} is inefficiency of the i$^{\text{th}}$ DMU at the first stage.

The J remaining equations model the transformation of J intermediate outputs to J final output in the second production stage. They can be specified as[36]

$$\ln\left(z_i^j\right) = \alpha_2^j + \beta_2^j \ln(y_i^j) + u_{2i}^j + v_{2i}^j, j = 1, \ldots J \tag{27}$$

where u_{2i}^j are inefficiency the j$^{\text{th}}$ division of the i$^{\text{th}}$ DMU at the second stage.

The overall inefficiency of the i$^{\text{th}}$ DMU is

$$u_i = u_{1i} + \sum_{j=1}^{J} \beta_1^j u_{2i}^j \tag{28}$$

It is noted that like the system of Eqs. (1) – (2), the system of $(J+1)$ Eqs. (26) – (27) has endogenous variables $\ln(z_i^j)$ on the right hand side of Eq. (26). Thus, 3SLS is used to estimate slope parameters.

Case 3: Intermediate outputs are exogenous (input and final outputs are endogenous)

Similar to case 1, we can extend the second equation, Eq. (25) to J equations for J divisions to exploit disaggregated data of intermediate and final outputs. Thus, the system consists of $(J+1)$ equations, specified as

Stage 1:

$$\ln(x_i) = \alpha_1 + \sum_{j=1}^{J} \beta_1^j \ln(z_i^j) + u_{1i} + v_{1i} \tag{29}$$

Stage 2:

$$\ln\left(y_i^j\right) = \alpha_2^j + \beta_2^j \ln\left(z_i^j\right) - u_{2i}^j + v_{2i}^j \tag{30}$$

$$(\text{Input}-\text{oriented})\text{overall inefficiency is} u_i^I = u_{1i} + \sum_{j=1}^{J} \frac{\beta_1^j}{\beta_2^j} u_{2i}^j. \tag{31}$$

The system of $(J+1)$ Eqs. (29) – (30) does not have the endogeneity[37].

[36] To make the system (26)-(27) be identified and the single-stage equation (and overall inefficiency) be derived by substituting $\ln(z_i^j)$ in Eq. (26) with the right hand side of Eq. (27), the system requires that the number of intermediate outputs is equal to the number of the second stage equations; each intermediate output is a dependent variable of each second stage equation.

[37] To get overall inefficiency expressed in Eq. (31), the system (29)-(30) has to have the number of intermediate outputs equal to the number of the second-stage equations and each second-stage equation obtains each intermediate output as an explanatory variable.

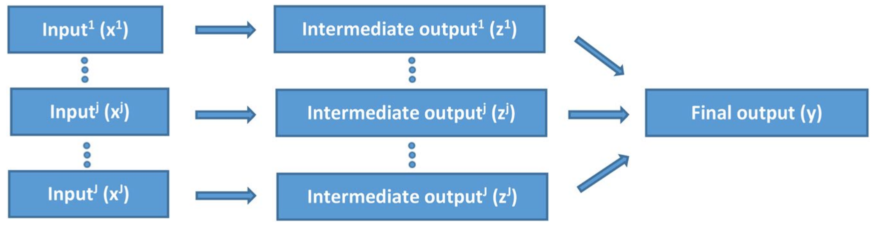


Fig. 5 The two-stage production of a multi-division DMU with aggregated final output

Fig. 6 The J-stage production process

- *Scenario 2*: data of input and intermediate output are at a division level, but the data of final outputs is only at a firm level; and firms' behavior is **case 2**.

Given that the data of inputs and intermediate outputs are available at a division level (Fig. 5), the first equation of the system (5) – (6) can transformed into J equations, specified as[38]

$$\ln\left(z_i^j\right) = \alpha_1^j + \sum\nolimits_{j=1}^{J} \beta_1^j \ln\left(x_i^j\right) - u_{1i}^j + v_{1i}^j \tag{32}$$

The second production stage is modelled by the following equation

$$\ln(y_i) = \alpha_2 + \sum\nolimits_{j=1}^{J} \beta_2^j \ln\left(z_i^j\right) - u_{2i} + v_{2i} \tag{33}$$

$$\text{The overall efficiency is } u_i = \sum\nolimits_{j=1}^{J} \beta_2^j u_{1i}^j + u_{2i} \tag{34}$$

Similar to the system of Eqs. (5) – (6), this system has an endogenous issue.

Appendix 3: Model extension for J-stage production process

Suppose there are J stage production process, described as Fig. 6. In the first stage, input (x) is used to produce the first intermediate output (z^1). The first intermediate output is used to produce the second intermediate output in the second stage and so on until the last intermediate output z^{J-1} is used to produce final output (y) in the last stage. The network SF model is a system of J equations.

Case 1: The final output is exogenous and input-oriented inefficiency

$$\begin{cases} Stage1 : \ln(x_i) = \alpha_1 + \beta_1 \ln\left(z_i^1\right) + u_{1i} + v_{1i} \\ \cdots \\ Stagej : \ln\left(z_i^{j-1}\right) = \alpha_j + \beta_j \ln\left(z_i^j\right) + u_{ji} + v_{ji} (1 < j < J) \\ \cdots \\ StageJ : \ln\left(z_i^{J-1}\right) = \alpha_J + \beta_J \ln(y_i) + u_{Ji} + v_{Ji} \end{cases}$$

Case 2: The input is exogenous and output-oriented inefficiency

$$\begin{cases} Stage1 : \ln\left(z_i^1\right) = \alpha_1 + \beta_1 \ln(x_i) - u_{1i} + v_{1i} \\ \cdots \\ Stagej : \ln\left(z_i^j\right) = \alpha_j + \beta_j \ln\left(z_i^{j-1}\right) - u_{ji} + v_{ji} (1 < j < J) \\ \cdots \\ StageJ : \ln(y_i) = \alpha_J + \beta_J \ln\left(z_i^{J-1}\right) - u_{Ji} + v_{Ji} \end{cases}$$

Case 3: The intermediate output z^J is exogenous $(1 \le j \le J)$, input-oriented inefficiency in the first J stage and input-oriented inefficiency in the rest

$$\begin{cases} Stage1 : \ln(x_i) = \alpha_1 + \beta_1 \ln\left(z_i^1\right) + u_{1i} + v_{1i} \\ \cdots \\ Stagej : \ln\left(z_i^{j-1}\right) = \alpha_j + \beta_j \ln\left(z_i^j\right) + u_{ji} + v_{ji} \\ Stage(j+1) : \ln\left(z_i^{j+1}\right) = \alpha_j + \beta_j \ln\left(z_i^j\right) - u_{(j+1)i} + v_{(j+1)i} \\ \cdots \\ StageJ : \ln(y) = \alpha_J + \beta_J \ln\left(z_i^{J-1}\right) - u_{Ji} + v_{Ji} \end{cases}$$

[38] Similar to the system (26)-(27), the system (32)-(33) requires that the number of intermediate outputs is equal to the number of the first-stage equations and each intermediate output is the dependent variable of each first-stage equation.

Appendix 4 The estimation steps for the systems specified in Subsection 4.1

Steps	System (1) – (2)	System (5) – (6)	System (9) – (10)
Step 1	3SLS	3SLS	SURE
Step 2	Third moments of residuals $\hat{\varepsilon}^*$ $\hat{m}_{31} = \sqrt{\frac{2}{\pi}}\left(\frac{4}{\pi}-1\right)\hat{\sigma}_{u1}^3$ $\hat{m}_{32} = \sqrt{\frac{2}{\pi}}\left(\frac{4}{\pi}-1\right)\hat{\sigma}_{u2}^3$	Third moments of residuals $\hat{\varepsilon}^*$ $\hat{m}_{31} = -\sqrt{\frac{2}{\pi}}\left(\frac{4}{\pi}-1\right)\hat{\sigma}_{u1}^3$ $\hat{m}_{32} = -\sqrt{\frac{2}{\pi}}\left(\frac{4}{\pi}-1\right)\hat{\sigma}_{u2}^3$	Third moments of residuals $\hat{\varepsilon}^*$ $\hat{m}_{31} = \sqrt{\frac{2}{\pi}}\left(\frac{4}{\pi}-1\right)\hat{\sigma}_{u1}^3$ $\hat{m}_{32} = -\sqrt{\frac{2}{\pi}}\left(\frac{4}{\pi}-1\right)\hat{\sigma}_{u2}^3$
	Second moments of residuals $\hat{\varepsilon}^*$ $\hat{m}_{21} = \left(1-\frac{2}{\pi}\right)\hat{\sigma}_{u1}^2 + \hat{\sigma}_{v1}^2$ and $\hat{m}_{22} = \left(1-\frac{2}{\pi}\right)\hat{\sigma}_{u2}^2 + \hat{\sigma}_{v2}^2$		
	Corrected intercepts: $\hat{\alpha}_1 = \hat{\alpha_1^*} - \sqrt{2/\pi}\hat{\sigma}_{u1}$ $\hat{\alpha}_2 = \hat{\alpha_2^*} - \sqrt{2/\pi}\hat{\sigma}_{u2}$	Corrected intercepts: $\hat{\alpha}_1 = \hat{\alpha_1^*} + \sqrt{2/\pi}\hat{\sigma}_{u1}$ $\hat{\alpha}_2 = \hat{\alpha_2^*} + \sqrt{2/\pi}\hat{\sigma}_{u2}$	Corrected intercepts: $\hat{\alpha}_1 = \hat{\alpha_1^*} - \sqrt{2/\pi}\hat{\sigma}_{u1}$ $\hat{\alpha}_2 = \hat{\alpha_2^*} + \sqrt{2/\pi}\hat{\sigma}_{u2}$
	Corrected residuals: $\hat{\varepsilon}_{1i} = \hat{\varepsilon}_{1i}^* + \sqrt{2/\pi}\hat{\sigma}_{u1}$ $\hat{\varepsilon}_{2i} = \hat{\varepsilon}_{2i}^* + \sqrt{2/\pi}\hat{\sigma}_{u2}$	Corrected residuals: $\hat{\varepsilon}_{1i} = \hat{\varepsilon}_{1i}^* - \sqrt{2/\pi}\hat{\sigma}_{u1}$ $\hat{\varepsilon}_{2i} = \hat{\varepsilon}_{2i}^* - \sqrt{2/\pi}\hat{\sigma}_{u2}$	Corrected residuals: $\hat{\varepsilon}_{1i} = \hat{\varepsilon}_{1i}^* + \sqrt{2/\pi}\hat{\sigma}_{u1}$ $\hat{\varepsilon}_{2i} = \hat{\varepsilon}_{2i}^* - \sqrt{2/\pi}\hat{\sigma}_{u2}$
Step 3	Stage inefficiency: $\hat{u}_{ki} = M\left(u_{ki}, \vert, \hat{\varepsilon}_{ki}\right)$ or $\hat{u}_{ki} = E\left(u_{ki}, \vert, \hat{\varepsilon}_{ki}\right)$		
	Overall efficiency: $\hat{u}_i = \hat{u}_{1i} + \hat{\beta}_1\hat{u}_{2i}$	Overall efficiency: $\hat{u}_i = \hat{\beta}_2\hat{u}_{1i} + \hat{u}_{2i}$	(Input-oriented) overall inefficiency: $\hat{u}_i^I = \hat{u}_{1i} + \frac{\hat{\beta}_1}{\hat{\beta}_2}\hat{u}_{2i}$

Appendix 5 Variable definition

Categories	Variables	Definition	Units	Sources
Costs	Maintenance cost	A sum of costs for (i) maintenance policy and planning, (ii) structural maintenance, and (iii) environmental, safety, and routine maintenance. It is noted that costs include revenue and capital expenditure but exclude new construction of roads.	£ thousand	Department for Levelling Up, Housing and Communities (DLUHC).
	Wage	Median hourly wages by regions in the civil engineering industry.	£ per hour	Office for National Statistics (ONS)
Intermediate outputs	Treatment on major roads	Lane-km of major roads received treatments	Km	Department for Transport (DfT)
	Treatment on minor roads	Lane-km of minor road received treatments	Km	DfT
Final outputs*	Major road condition	Proportion of major roads without maintenance needs.	Km	DfT
	Minor road condition	Kilometers of minor roads without maintenance needs.	Km	DfT
Control variables	Land area	land area of each authority	Hectare	ONS
	Length of major roads	Lane-km of major roads	Km	DfT
	Length of minor roads	Lane-km of minor road	Km	DfT
	Traffic density in major roads	A ratio of thousand vehicle kilometers travelled on major roads to major roads' length.	1000 vehicles	DfT
	Traffic density in minor roads	A ratio of thousand vehicle kilometers travelled on minor roads to minor roads' length.	1000 vehicles	DfT

*In the model we use change of road condition as final output rather than road condition

Appendix 6 Estimated sigma u, sigma v and intercepts of the Network SF model by different methods

Parameters	Stage	Method of moments	Copula-based MLE
Sigma u	Stage 1	0.76	0.74
	Stage 2 - Major	0.50	0.35
	Stage 2 - Minor	0.38	0.21
Sigma v	Stage 1	0.49	0.50
	Stage 2 - Major	0.60	0.64
	Stage 2 - Minor	0.57	0.60
Intercept	Stage 1	7.99	8.01
	Stage 2 - Major	−6.69	−6.57
	Stage 2 - Minor	−5.10	−4.97

Appendix 7: Stata package for network SFA

The Stata package "networkSFA" is available for free download at https://github.com/trangtvh/networkSFA.

To install the package, enter the following commands into Stata. (If you already have the github command installed, the first command is not needed):

net install github, from("https://haghish.github.io/github/")

github install trangtvh/networkSFA

After the installation, enter the following commands for instruction on how to use the package:

help networkSFA

help networkSFA_postestimation

Acknowledgements The authors thank the editor, the associate editor and two anonymous referees for their helpful comments and recommendations. The author acknowledge funding from the CQC Efficiency Network (see http://www.nhtnetwork.org/cqc-efficiency-network/home/).

Author contributions VHT Tran was responsible for data collection, estimating results, and writing the first draft.PE Wheat and AD Stead refined the draft and finalised the final manuscripts.All authors contributed to building the model and the estimation method proposed in the manuscripts and reviewing the manuscripts.

Data availability The dataset used in the empirical example of this paper is not publicly available but is available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors declare no competing interests.

References

Amemiya T (1977) The maximum likelihood and the nonlinear three-stage least squares estimator in the general nonlinear simultaneous equation model. Econometrica 45(4):955–968

Amsler C, Prokhorov A, Schmidt P (2014) Using copulas to model time dependence in stochastic frontier models. Econom Rev 33(5-6):497–522

Andor M, Hesse F (2014) The stoned age: the departure into a new era of efficiency analysis? a monte carlo comparison of stoned and the 'Oldies' (SFA and DEA). J Prod Anal 41(1):85–109

Andor M, Parmeter C (2017) Pseudolikelihood estimation of the stochastic frontier model. Appl Econ 49(55):5651–5661

Charnes A, Cooper WW, Golany B, Halek R, Klopp G, Schmitz E, Thomas D (1986) Two-phase data envelopment analysis approaches to policy evaluation and management of army recruiting activities: Tradeoffs between joint services and army advertising. Research Report CCS #532, Center for Cybernetic Studies. University of Texas-Austin, Austin

Chen Y, Cook WD, Li N, Zhu J (2009) Additive efficiency decomposition in two-stage DEA. Eur J Oper Res 196(3):1170–1176

Chiou YC, Lan LW, Yen BTH (2010) A joint measurement of efficiency and effectiveness for non-storable commodities: Integrated data envelopment analysis approaches. Eur J Oper Res 201(2):477–489

Coelli TJ (1995) Estimators and hypothesis tests for a stochastic frontier function: a monte carlo analysis. J Prod Anal 6(4):247–268

Cook W, Liang L, Zhu J (2010) Measuring performance of two-stage network structures by DEA: A review and future perspective. Omega 38:423–430

Despotis DK, Sotiros D, Koronakos G (2016a) Composition versus decomposition in two-stage network DEA: a reverse approach. J Prod Anal 45:71–87

Despotis DK, Sotiros D, Koronakos G (2016b) A network DEA approach for series multi-stage processes. Omega 61:35–48

Fallah-Fini S, Triantis K, Garza JM, Seaver WL (2012) Measuring the efficiency of highway maintenance contracting strategies: A bootstrapped non-parametric meta-frontier approach. Eur J Oper Res 219(1):134–145

Fallah-Fini S, Triantis K, Rahmandad H, Garza JM (2015) Measuring dynamic efficiency of highway maintenance operations. Omega 50:18–28

Färe R, Grosskopf S (1996) Intertemporal production frontiers: With dynamic DEA. Kluwer Academic Publishers, Boston

Fukuyama H, Matousek R (2017) Modelling bank performance: A network DEA approach. Eur J Oper Res 259(2):721–732

Greene W (2012) Econometric analysis, 7th edn. Prentice Hall, Upper Saddle River

Halkos GE, Tzeremes NG, Kourtzidis S (2014) A unified classification of two-stage DEA models. Surv Oper Res Manag Sci 19:1–16

Hausman JA (1983) Chapter 7 Specification and estimation of simultaneous equation models. In: Griliches Z, Intriligator MD (eds), Handbook of econometrics. vol 1. Elsevier, p 391–448

Heckman JJ, Robb RJ (1985) Alternative methods for evaluating the impact of interventions: An overview. J Econom 30(1-2):239–267

Huang TH, Chen KC, Lin CI (2018) An extension from network DEA to copula-based network SFA: Evidence from the U.S. commercial banks in 2009. Q Rev Econ Finance 67:51–62

Huang TH, Lin CI, Chen KC (2017) Evaluating efficiencies of Chinese commercial banks in the context of stochastic multistage technologies. Pac Basin Finance J 41:93–110

Jondrow J, Knox Lovell CA, Materov IS, Schmidt P (1982) On the estimation of technical inefficiency in the stochastic frontier production function model. J Econom 19(2–3):233–238

Kalb A (2014) What determines local governments' Cost-efficiency? The case of road maintenance. Reg Stud 48(9):1483–1498

Kao C (2014) Network data envelopment analysis: A review. Eur J Oper Res 239:1–16

Kao C, Hwang SN (2008) Efficiency decomposition in two-stage data envelopment analysis: An application to non-life insurance companies in Taiwan. Eur J Oper Res 185(1):418–429

Kumbhakar SC, Tsionas EG (2006) Estimation of stochastic frontier production functions with input-oriented technical efficiency. J Econom 133(1):71–96

Lai H, Huang CJ (2013) Maximum likelihood estimation of seemingly unrelated stochastic frontier regressions. J Prod Anal 40(1):1–14

Lan LW, Lin ETJ (2006) Performance measurement for railway transport: Stochastic distance functions with inefficiency and ineffectiveness effects. J Transp Econ Policy 40(3):383–408

Lin YH, Hong CF (2020) Efficiency and effectiveness of airline companies in Taiwan and Mainland China. Asia Pac Manag Rev 25(1):13–22

Liu D (2016) Measuring aeronautical service efficiency and commercial service efficiency of East Asia airport companies: An application of Network Data Envelopment Analysis. J Air Transp Manage 52:11–22

Matthews K (2013) Risk management and managerial efficiency in Chinese banks: A network DEA framework. Omega 41(2):207–215

Office of Rail and Road (ORR) (2023) Annual assessment of National Highways' performance - April 2022 to March 2023. Avaliable at https://www.orr.gov.uk/monitoring-and-regulation/roads-monitoring/annual-assessment-national-highways

Olson JA, Schmidt P, Waldman DA (1980) A Monte Carlo study of estimators of stochastic frontier production functions. J Econom 13:67–82

Ozbek ME, Garza JM, Triantis K (2012) Efficiency measurement of the maintenance of paved lanes using data envelopment analysis. Constr Manag Econ 30(11):995–1009

Peyrache A, Silva MC (2022). Efficiency and productivity analysis from a system perspective: Historical overview. In: Chotikapanich D, Rambaldi AN, Rohde N (eds), Advances in economic measurement: A volume in honour of DS Prasada Rao, Springer, p 173–230

Rouse P, Chiu T (2009) Towards optimal life cycle management in a road maintenance setting using DEA. Eur J Oper Res 196(2):672–681

Seiford LM, Zhu J (1999) Profitability and marketability of the top 55 U.S. commercial banks. Manage Sci 45:1270–1288

Sjögren S, Söderberg M (2011) Productivity of airline carriers and its relation to deregulation, privatisation and membership in strategic alliances. Transp Res E Logist Transp Rev 47(2):228–237

Stead AD, Wheat P (2020) The case for the use of multiple imputation missing data methods in stochastic frontier analysis with illustration using English local highway data. Eur J Oper Res 280(1):59–77

Tavassoli M, Faramarzi GR, Saen RF (2014) Efficiency and effectiveness in airline performance using a SBM-NDEA model in the presence of shared input. J Air Transp Manage 34:146–153

The Department of Transport (2012) Guidance on road classification and the primary route network. Available at: https://www.gov.uk/government/publications/guidance-on-road-classification-and-the-primary-route-network/guidance-on-road-classification-and-the-primary-route-network. Accessed 26 June 2025

Tsay WJ, Huang CJ, Fu TT, Ho IL (2013) A simple closed-form approximation for the cumulative distribution function of the composite error of stochastic frontier models. J Prod Anal 39(3):259–269

Wang CH, Gopal RD, Zionts S (1997) Use of data envelopment analysis in assessing information technology impact on firm performance. Ann Oper Res 73:191–213

Wheat P (2017) Scale, quality and efficiency in road maintenance: Evidence for English local authorities. Transp Policy 59:46–53

Wheat P, Stead AD, Greene WH (2019) Robust stochastic frontier analysis: A student's t -half normal model with application to highways maintenance costs in England. J Prod Anal 51(1):21–38

Wooldridge JM (2015) Control Function Methods in Applied Econometrics. J Hum Resour 50(2):420–445

Yarmukhamedov S, Smith ASJ, Thiebaud J (2020) Competitive tendering, ownership and cost efficiency in road maintenance services in Sweden: A panel data analysis. Transp Res A Policy Pract 136:194–204

Yu MM, Lin ETJ (2008) Efficiency and effectiveness in railway performance using a multi-activity network DEA model. Omega 36(6):1005–1017

Zellner A, Theil H (1962) Three-stage least squares: simultaneous estimation of simultaneous equations. Econometrica 30(1):54–78