Journal of Productivity Analysis (2024) 62:217–238
https://doi.org/10.1007/s11123-024-00731-5

The inefficiency of courts of justice: industry structure, capacity and misallocation

Antonio Peyrache[1] · Angelo Zago[2]

Accepted: 7 July 2024 / Published online: 25 July 2024

Abstract
In this paper, we introduce an optimization model to quantify the trade-off between resource capacity utilization and disposition time for the caseload of courts of justice. The optimization model takes into account the impact of an increase in demand that may arise when disposition time is reduced. We employ the model to measure the impact of various policy reform scenarios on the length of trials, both at the court and system level. We do so by taking into account the potential reallocation of resources, using the population of Italian courts of justice over the 2005–2012 period. Our results show that if all policy scenarios we discuss were to be implemented, the average length of trials for civil cases would be more than halved, from the current 15.5 months to about 7 months. Implementing best practices, the single most effective policy would be equivalent to a 25% increase in the number of judges (which would otherwise cost around 100 million euros per year).

Keywords Data Envelopment Analysis · Disposition times · Optimal resource allocation · Demand feedback effects

JEL classification C44 · L23 · L38 · C23

1 Introduction

Public services (such as health, education, justice, etc.) account for a large proportion of economic activity. These services operate without a clear efficient market and price mechanism and are therefore potentially prone to inefficiencies in the use of resources or to excessive waiting times. In this paper, we introduce an optimization model that quantifies the trade-off between resource capacity utilization and service time for decision-making units (DMU) that operate multiple services. For example, a hospital manages various medical specialities (cardiology, radiology, etc.) within the same administrative unit and these services may share some resources. The model also takes into account the potential additional caseload that may arise when service time is reduced (and the opportunity cost of using the service is reduced) via a predictive model.

We focus our empirical analysis on the judicial system. Judicial systems (and the rule of law) have a very important role in securing property rights and enforcing contracts, thus affecting economic behavior, investment choices and economic growth (Aldashev, 2009). Furthermore, the judiciary is one sector of the economy where the market system cannot work, given the absence of a functioning output price mechanism that could penalize inefficient courts. In this paper, we undertake a quantitative analysis of courts of justice by focusing on supply policies designed to increase the efficiency of the system. While most of the studies we are aware of (illustrated below) consider the demand for justice and related policies, we provide a model that can account for inefficiencies arising from the supply side and consider the impact of different policies to improve the efficiency of the Italian judicial system.

The literature on the efficiency of courts of justice is relatively limited. [1]Lewin et al. (1982) is probably the first

✉ Angelo Zago
angelo.zago@univr.it

[1] School of Economics, University of Queensland, St. Lucia, QLD, Australia

[2] Dipartimento di Scienze Economiche, Università degli Studi di Verona, Verona, Italy

[1] As documented by Pereira et al. (2023), out of 6100 publications in the field of Law and Economics, only 44 are related to judicial efficiency in Europe, many dealing with the Italian courts.

study, dealing with the superior (criminal) court of North Carolina in 1976 using the so-called CCR input-oriented measure developed by Charnes et al. (1978). Kittelsen and Forsund (1992) investigate the technical efficiency (with both input- and output-oriented CCR and the BCC measure developed by Banker et al. 1984) and calculate a Malmquist index of productivity to analyze Norwegian courts from 1983 to 1988. Tulkens (1993) used FDH to investigate Belgian courts during 1983–1985. Pedraja-Chaparro and Salinas-Jimenez (1996) estimated input-oriented CCR and BCC measures of Spanish courts in 1991.[2]

Most if not all of the literature on the efficiency of courts considers variants of the DEA methodology, together with resource use (judges and other staff, mainly) on the input side and defined cases on the output side (among recent contributions along these lines see, e.g., Santos and Amado 2014; Silva 2018; Chen et al. 2021; Kerstens and Xiaoqing 2022). However, a measure of performance often used by practitioners is the average length of trials, seen among other things as a tool to identify problematic countries or courts within a country (see CEPEJ 2016). As explained in a recent OECD report, "the focus on length is motivated not only by the importance of a timely resolution of disputes for the correct functioning of the economy, but also by the fact that a reasonable trial length is a necessary (though not a sufficient) condition for good performance in other dimensions [..] Also, as emphasized by the adage *justice delayed is justice denied*, timeliness is a prerequisite for achieving justice. Moreover, the length of trials is also generally associated with other crucial measures of performance such as confidence in the justice system" (Palumbo et al. 2013: p. 9).

Although completion time[3] is pivotal in evaluating the justice system and other public services, looking only at this single KPI without considering resource use is not fully informative and may be misleading. Therefore, the first contribution of our paper is to take into account resource use in an optimization model that explicitly introduces the trade-off between the length of trial and resource use. We thus improve on the standard practice of looking only at trial length, since our model enables us to understand what causes delays in justice from a supply side (resource) perspective.

Another feature of our modeling strategy accounts for the fact that a shortening of disposition time may lead to an increase in demand. This is due to the fact that processing time may act as a rationing-by-waiting mechanism. An increase in the demand for justice may offset some of the benefits of an increased processing speed obtained by the supply side. We thus estimate a demand function and incorporate it into our equilibrium outcome in order to account for the *demand feedback effect* on trial length. Therefore our optimization program allows us to account for resource use, disposition times, and potential demand feedback effects that may increase the number of incoming cases when trial length is reduced. The optimization approach can be used to study the (steady state) performance of the justice sector as a whole and improve on both the analysis of trial length and the standard measures of partial productivity (the number of completed cases per judge).

Finally, in order to account for resource use, we consider a production frontier for the courts of justice where the number of pending cases is a *variable* input and the units of personnel are a *fixed* production factor (a capacity input). For a given quantity of human resources (judges and administrative staff), when the number of pending cases increases, the number of completed cases first increases, then reaches a maximum and finally decreases due to congestion. For this reason, the model can accommodate both variable returns to scale and production congestion. Congestion may cause additional problems especially in courts with a large stock of pending cases.[4]

Such an approach allows us to relate the time needed to complete cases to the possible causes of excessive trial length, enabling us to make policy suggestions targeting the sources of inefficiency (and their geographical distribution). In particular, we consider four supply policies that were either implemented or discussed in the Italian case: the introduction of best practices at the court level; the break-up of large courts (to avoid diseconomies of scale); the increase in personnel; and the optimal reallocation of personnel to

[2] Earlier studies investigating the Italian judicial system include Marchesi (2003, 2008), who uses an input requirement function (mainly because on the input side only information on judges is available) with data on Italian courts for 1996, 2001, and 2006. Marselli and Vannini (2004), with input-oriented CCR and BCC measures, investigated the 29 Appeal Court Districts in 2002. Lastly, Ricolfi (2009) uses a linear production function and data for 2005. For more recent contributions on the efficiency of Italian courts see the literature cited in Pereira et al. (2023).

[3] *Completion time* is the effective duration of a trial, calculated as the difference between the date of registration and the date on which the ruling or settlement decision is published. The *disposition time*, on the other hand, provides an estimate of the average foreseeable time for defining proceedings by comparing the stock of pending cases at the end of the year with the flow of proceedings defined during the year. The duration is therefore approximated by the time necessary to exhaust the ongoing proceedings, assuming that the capacity to dispose of the trials remains constant and there are no new proceedings registered. While completion time is more accurate, the lack of data often leads to using disposition time as its proxy (see, e.g., CEPEJ 2016). Our data enable us to consider disposition time. However, we sometimes use it interchangeably with processing time and trial length.

[4] Congestion may be due, for instance, to *task juggling* (Coviello et al. 2014). The first paper to acknowledge congestion problems in courts is probably Buscaglia and Dakolias (1999), but to the best of our knowledge few other papers have dealt with it, e.g., Dimitrova-Grajzla et al. (2012), Coviello et al. (2015), and Bray et al. (2016).

courts. We consider the effect of these policies both on the average disposition time of the system and on the distribution of disposition times across the different courts.

1.1 Why the Italian case?

Apart from being useful to illustrate our methodology, the Italian court system is interesting because of its poor performance and its heterogeneity. Italian courts are among the most inefficient among OECD countries in terms of trial length. The *average disposition time* for a standard commercial case in 2016 was 1120 days in Italy, against 553 in OECD countries (regional average), 395 in France, 499 in Germany and 510 in Spain. According to the World Bank (Doing Business, 2020 edition), Italy ranks 122nd out of 190 countries in terms of enforcing contracts, compared to Germany (13th), France (16th), Spain (26th), and UK (34th). Moreover, the average trial length is quite different over Italian regions, with lengthier processes and larger stocks of pending cases in the South. However, given that southern courts are provided with more human resources, it is important "to establish whether and to what extent the larger stock of pending cases is due to lack of resources or to their lower productivity" (Carmignani and Giacomelli 2009: 21).

Thus it is not surprising that the Italian justice system has been investigated quite extensively. In recent years, there has been a lively discussion of the possible causes of these inefficiencies and, in particular, of *pathological demand* effects (Marchesi 2003), according to which higher litigation rates are the result of lengthy trials. Delays in delivering justice could lead some economic agents (households, workers and firms) to exploit these inefficiencies by strategically postponing their contractual obligations to other parties, and this is more likely to happen the wider the gap between legal [5] and market interest rates (see, e.g., Marchesi 2003; Felli et al. 2008; Padrini et al. 2009). Other theories point to *supplier-induced demand*, (see, e.g., Carmignani and Giacomelli 2010 and Buonanno and Galizzi 2014), according to which the combination of the increase in the number of lawyers leads to excessive litigation. However, the empirical evidence regarding these possible demand-side causes is rather ambiguous, and numerous studies call for a complementary supply-side analysis (see, e.g., Bianco and Palumbo 2007; Felli et al. 2008). Indeed, albeit a country with one of the highest litigation rates, Italy is given as the example where "there is scope for improvements also on the supply side, for instance expanding the use of case-flow management techniques" (Palumbo et al. 2013: 45), a policy aimed at introducing best (management) practices.

To empirically implement our model, we collected data for all Italian courts (165) for the 2005–2012 period, [6] taking advantage of data now publicly available and collecting additional data from other sources. Overall, we find that technical (best practices), size (break-ups) and reallocation inefficiencies are the major issues at the industry level. Given these findings, we argue that the most effective policy would be the introduction of best practices, which would have effects throughout the system, including in the inefficient courts of southern Italy. Another effective policy might be to increase personnel, although its cost-effectiveness and hence feasibility might be questioned.

Section 2 illustrates the computational models, section 3 the data, and section 4 the empirical results. The final section concludes with some suggestions for further research. After a brief review of the significant literature, the Appendix presents a market justice model and additional results.

2 Methodology

We consider a service industry composed of $j = 1, \ldots, J$ decision-making units (DMU). Each DMU provides $p = 1, \ldots, P$ services. On each service line, new incoming cases arrive and queue together with the existing pending cases of that particular service, waiting to be processed. The P services will process cases in the queue by using economic resources (inputs). At any given point in time, service p of DMU j will face w_{pj} pending cases that are queuing, waiting to be serviced (processed). To process cases the DMU will allocate Q available inputs to the different services. For some inputs the allocation to individual processes is not observed, only the total input use at the DMU level is. There are M inputs whose allocation is not observed, leaving $N = Q - M$ inputs for which the allocation is observed. Clearly, if $N = 0$ we do not observe the allocation of the inputs to the different services and if $N = Q$ we observe all allocations. The data for DMU j on the inputs for which we observe the allocation are stored in a $N \times 1$ vector $\mathbf{x}_{pj}$: this is the amount of inputs used by service p in DMU j. The data for DMU j on the inputs for which we do not observe the allocation are stored in a $M \times 1$ vector $\mathbf{z}_j$, giving the total amount used by DMU j (without the allocation to the separate P services).

[5] The legal interest rate is applied to borrowers for delayed payments of their debts when their case goes to court. It is determined annually by the Italian Ministry of Finance and is usually equal to the interest rate paid on Government bonds, i.e., it is normally lower than the commercial bank rate.

[6] Before the change in court geography introduced at the end of 2012 by the Monti Government.

We also observe the number of cases y_{pj} that each service p in each DMU j has processed in the given reference time period. Additionally, the number of observed processed cases of service p may differ from the number of incoming cases. The number of incoming cases for service p in DMU j is calculated using a predictive model (see the online Appendix) and the relevant coefficients for such a prediction are stored in the set of coefficients v_{pj}, e_{pj}, i.e., one set of coefficients for each service of each DMU.

Figure 1 provides a graphic representation of how the model is applied to the courts. The figure represents a generic court of justice j. There are two main services provided to the public by the court of justice: process 1 describes civil cases and process 2 criminal cases. w^t_{1j} is the number of pending civil cases. In the given reference period t, y^t_{1j} is the number of civil cases that are processed by the court. During the reference period a number of additional incoming civil cases will be added to the queue and this means that in period $t+1$ the number of pending cases can be different from w^t_{1j} and are equal to $w^{t+1}_{1j} = w^t_{1j} - y^t_{1j} + inc^t_{1j}$, where inc^t_{1j} is the number of incoming cases. Clearly if $w^t_{1j} = w^{t+1}_{1j} = w_{1j}$ the queue is in steady-state and this means that the number of processed cases is equal to the number of incoming cases and this quantity is also time invariant: $y^t_{1j} = inc_{1j} = y_{1j}$. The expected disposition time for a new incoming case at the steady-state is equal to w_{1j}/y_{1j} and this quantity is clearly affected by the quantity of economic resources devoted to that particular service.[7]

A similar description holds for this figure in the case of criminal cases (service 2). The flow of cases is processed by a stock of inputs represented in this case by x and z. For example, x_{1j} is the number of judges assigned exclusively to civil cases and x_{2j} is the number of judges assigned exclusively to the processing of criminal cases. z_j is the total number of judges working in DMU j for which we do not observe the allocation to the two types of cases. These judges may well specialize only in criminal or civil cases, but we do not observe their allocation. It may as well be the case that these judges can deal with both criminal and civil cases and they are allocated based on the needs of the court. But, again, we are not able to observe the amount of time that they spend on each activity. It is clear from this discussion that the same

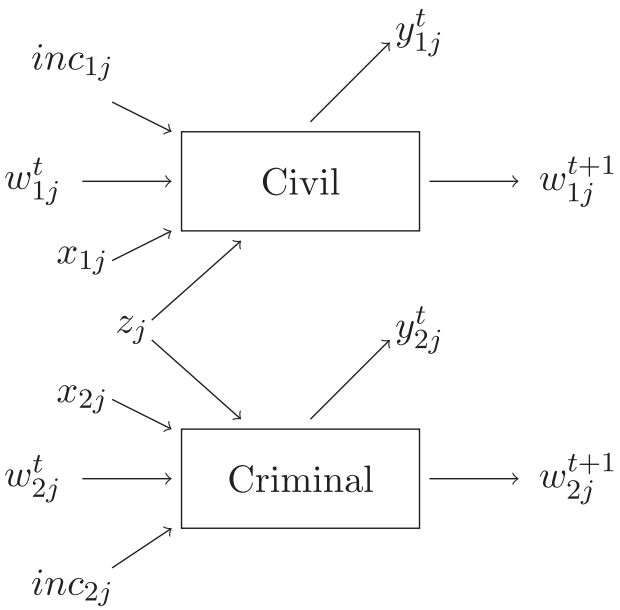

Fig. 1 Production Network for the Courts of Justice

data structure apply to a number of other public services such as health and education. All the mathematical programs illustrated below apply in these cases as well, as long as the data structure follows the description just given. For example, if the DMU were a hospital, the different service processes would represent different medical specialties (such as cardiology, radiology, etc.), the cases processed would correspond to the number of patients in each specialty, and the inputs would be doctors, nurses, etc.

The goal of the system is to service incoming cases in the shortest possible time given the number of cases pending in the queue and the economic resources the system is given. We shall return to this point later. Here we would like to stress that for this public service it is possible to use the above data to model the service as a production model. The production possibilities associated with this model can be made operational using a modified version of the approach presented in Podinovski (2021). This is useful because it provides a quantification of the trade-off between the use of economic resources and the speed with which cases are processed. Within this framework, we are, as a first step, interested in measuring the efficiency with which the operation of DMU k is conducted. Unit k can be either a DMU from the dataset or a hypothetical unit. We posit the following program to measure the efficiency of DMU k[8]

$$\begin{aligned} &\max \theta_k \\ st \quad &\sum_j \lambda^k_{pj} \mathbf{x}_{pj} \leq \mathbf{x}_{pk}, \ \forall p \end{aligned} \tag{1a}$$

[7] We took this definition from "Introduction to Operations Research" (Hillier, Lieberman, Tenth Edition). On page 736 when they introduce Queuing theory terminology they define the waiting time as the ratio of the queue length to the number of new arrivals in the unit of time. In our context the queue length is the number of pending cases and the new arrivals are the number of incoming cases. We provide a material balance condition that states that in the steady state (i.e., a situation where the number of pending cases does not change over time; as is the case for the Italian system), the number of incoming cases is the same as the number of processed cases. Therefore in steady-state, our measure of pending cases over processed cases is indeed the waiting time.

[8] In the programs below, Greek letters represent decision variables and Latin letters data; given this arrangement we do not need to specify the list of decision variables under the *max* or *min* operators.

$$\sum_j \delta_j^k \mathbf{z}_j \leq \mathbf{z}_k \tag{1b}$$

$$\sum_j \lambda_{pj}^k w_{pj} = w_{pk}, \ \forall p \tag{1c}$$

$$\sum_j \lambda_{pj}^k y_{pj} \geq y_{pk}\theta_k, \ \forall p \tag{1d}$$

$$\sum_j \lambda_{pj}^k = \sigma_k, \ \forall p \tag{1e}$$

$$\delta_j^k \geq \lambda_{pj}^k, \ \forall p, j \tag{1f}$$

This program seeks the maximum expansion of the processed cases y_{pk} compatible with constraints on the use of resources. Constraint (1a) and (1b) state that input use cannot exceed the observed inputs of DMU k. Constraint (1c) states that the chosen linear combination of DMUs must return the same number of pending cases of DMU k. The equality in this constraint means that pending cases are allowed to be a congesting factor (see Wei and Yan, 2004). Constraint (1d) looks for the maximum expansion of the observed processed output compatible with the specified production set. Constraint (1e) can be used to model scale economies and determine the optimal size of the DMU. Constraint (1f) is discussed in detail in Podinovski (2021) and reflects the fact that the allocation of the inputs $\mathbf{z}$ is not observed. In this program we treat pending cases as a potentially congesting factor: they enter the program with an equality constraint and this means that they behave as a variable input. This also means that the model does not rule out the possibility that an excessive number of pending cases could generate congestion.

We solve this program for each DMU k in the dataset and under two different assumptions for parameter σ_k. First, we assume that $\sigma_k = 1$ and denote the optimum value obtained for the objective function as $A_k = \theta_k^*$. We assume that $\sigma_k \geq 1$ and denote the optimum value obtained for σ as $S_k = \sigma_k^*$. Hence, from this first program we obtain a pair of values (A_k, S_k) for each DMU k. The value A_k represents the efficiency with which the DMU is operating and the value S_k the optimal scale of operation for that particular DMU. Indeed, following Peyrache and Zago (2016), S_k provides a benchmark to determine if a DMU should be split into a number of smaller units in order to avoid the negative effects of decreasing returns to scale.

We should also stress that the fact that we have an equality constraint on the number of pending cases, together with an inequality constraint on the inputs, means that this program empirically reflects what has been called the law of variable proportions (see Svensson and Färe, 1980): given the number of judges and other inputs (our capacity measure) when the number of pending cases increases, the number of processed cases increases at first, peaks and then decreases (due to congestion). The efficiency score A_k measures the distance from the frontier in terms of the additional number of pending cases that could be processed when the court is benchmarked against other courts of similar size. Therefore it measures the efficiency of the DMU compared to other DMUs representing the industry best practices.

Neither the observed combination (y_k, w_k) nor the efficient combination $(y_k A_k, w_k)$ are necessarily steady-state levels of service provision, since the average disposition time resulting from these quantities may be incompatible with the predicted values of the demand feedback effects. In order to define a steady-state outcome that keeps the efficiency level of the court constant at the observed level A_k, we use the following program where the relationship derived from the predictive model of the demand feedback effect is explicitly included:

$$\begin{aligned} &\min \sum_p \pi_{pk} \\ &st \sum_j \lambda_{pj}^k \mathbf{x}_{pj} \leq \mathbf{x}_{pk}, \ \forall p \end{aligned} \tag{2a}$$

$$\sum_j \delta_j^k \mathbf{z}_j \leq \mathbf{z}_k \tag{2b}$$

$$\sum_j \lambda_{pj}^k w_{pj} = \pi_{pk}, \ \forall p \tag{2c}$$

$$\sum_j \lambda_{pj}^k y_{pj} \geq A_k\left(v_{pk} + e_{pk}\pi_{pk}\right), \ \forall p \tag{2d}$$

$$\sum_j \lambda_{pj}^k = S_k, \ \forall p \tag{2e}$$

$$\delta_j^k \geq \lambda_{pj}^k, \ \forall p, j \tag{2f}$$

where we now let the number of pending cases π_{pk} be a decision variable. The demand feedback effect means that the number of processed cases is $\tau_{pk} = v_{pk} + e_{pk}\pi_{pk}$, which explains the meaning of constraint (2d), where the right-hand side is determined using the predicted demand. This demand effect (for $e_{pk} \leq 0$) states that when the processing time of a service is shortened there is an increase in the number of incoming cases (since this reduces the opportunity cost of using the service). If the number of pending cases π_{pk} becomes very small, the number of incoming cases is given by the data coefficients v_{pk} which are obtained via our predictive model. By minimizing the number of pending cases, this program implicitly seeks to

minimize disposition time given feasibility production constraints and taking into account the feedback effect that a shorter time has on the number of incoming cases.[9] Since the level of efficiency A_k is kept constant at the level determined by program (1), the disposition time as determined by this program is a steady-state disposition time for the given level of efficiency A_k. At the steady-state the average disposition time of the court does not change over time and the average number of incoming cases is equal to the average number of processed cases. In other words, this routine returns the *steady-state* average disposition time for efficiency level A_k as opposed to the *observed* average disposition time w_{pk}/y_{pk}.

2.1 Full efficiency

We can now show what happens to the steady-state completion time when we introduce best practices and increase the level of efficiency to $A_k = 1$. This involves solving the same optimization program (2) with the full rather than observed efficiency values. The comparison of the steady-state disposition time at full efficiency vs. the steady-state disposition time at a given level of efficiency produces a measure of the efficiency of the court in terms of disposition times.

It should be noted that this concept of efficiency is a steady-state notion, since it includes a constraint that takes into account the behavior of the demand feedback effect when the average disposition time changes. On the contrary, A_k is a measure of optimality irrespective of the level of demand for the service. In other words, A_k is potential rather than something that can be realized, and should be used to assess if a court is on the frontier or in the interior of the set. It should also be noted that in general the observed time (at the given level of efficiency A_k) may not comply with our steady-state concept. Our measure of time efficiency compares two alternative steady states: one with the observed level of inefficiency and the other with the court lying at the frontier.

2.2 The optimal size of courts

Technical inefficiency is far from the only component of inefficiency in the system. Two further types are explored below and in the following section: inefficiencies from diseconomies of scale and inefficiency from the non-optimal allocation of resources to the various DMUs and the various services.

[9] Note that if the number of services P is small enough, we could introduce a set of weights b_{pk} in the objective function and use them in a critical line algorithm to search for all possible trade-offs between the processing times of the different services.

To account for diseconomies of scale we solve program (2) using two values of S_k. First, we do so by using the value $S_k = 1$, therefore comparing the DMU to other DMUs of similar size. Second, we use the value of S_k computed in program (1). This second option compares the DMU to DMUs operating at a different (smaller) scale. If the two solutions differ, then this is the effect diseconomies of scale have on disposition time. In this case the solution is to split the unit under analysis into a number of smaller units, in other words the solution is to implement a break-up. If diseconomies of scale prevail, then breaking-up a large DMU into smaller ones increases the processing ability of the DMU and therefore shortens processing time.

We use a standard DEA-VRS model, which means that increasing returns are not automatically ruled out. In fact we find that many of the small courts operate in the increasing returns region. However, from a systemic point of view this is not the main contributor of inefficiency in the Italian system. In fact, large courts of justice (with over 50 judges) account for the great majority of resource use for the system. Just to provide some simple statistics, the largest 15% of courts (more than 50 judges) employ around 50% of judges; and the 5 largest courts (located in the 5 largest cities, with over 140 judges), while representing only 3% of the overall number of courts, employ 25% of the total number of judges. On the contrary, small courts (say fewer than 25 judges), account for more than 60% of the total number of courts, but only employ around 25% of the total number of judges. Thus, although some increasing returns may be relevant in this context, the overwhelming evidence is that the main problem is with the break-up of large courts. To be sure, we do not suggest that increasing returns for small courts are unimportant (they are accounted for in the reallocation component), but do suggest that break-ups are more important. Doubtless, it would be interesting from a methodological perspective to look at increasing returns to scale, but this is outside the scope of this paper, focused on the efficiency of the (Italian) system.

2.3 Reallocation efficiency

Another supply policy scenario we consider is the reallocation of resources across courts. This may enhance the efficiency of the system by moving inputs from courts which have very fast disposition times to courts with much slower disposition times; or, similarly, it could consider gains obtainable when the ratio of inputs is not optimal or the quantity of some inputs is found not to constrain service provision, i.e., there is excess capacity. The reallocation problem for the system as a whole can be written as in program (3), which can be solved in a single step for all DMUs. This problem looks at the minimization of the overall number of pending cases (therefore

minimization of disposition time), given demand constraints for each DMU as modeled via the predictive model.

The goal of the system is to service incoming cases in the shortest possible time given the number of cases pending in the queue. Program (3) embeds this idea. The objective function has the same interpretation as routine (2) with the difference that now pending cases are summed across all the DMUs k under evaluation:

$$\min \sum_k \sum_p \pi_{pk}$$

$$st \quad \sum_j \lambda_{pj}^k \mathbf{x}_{pj} \leq \mathbf{x}_{pk} + \boldsymbol{\mu}_{pk} - \sum_i \sum_q \boldsymbol{\alpha}_{ki}^{pq} + \sum_i \sum_q \boldsymbol{\alpha}_{ik}^{qp}, \ \forall p,k \tag{3a}$$

$$\sum_k \sum_p \left[\sum_i \sum_q \boldsymbol{\alpha}_{ki}^{pq} - \sum_i \sum_q \boldsymbol{\alpha}_{ik}^{qp} \right] \leq 0 \tag{3b}$$

$$\sum_j \delta_j^k \mathbf{z}_j \leq \mathbf{z}_k + \boldsymbol{\eta}_k - \sum_i \boldsymbol{\gamma}_{ki} + \sum_i \boldsymbol{\gamma}_{ik}, \ \forall k \tag{3c}$$

$$\sum_k \left[\sum_i \boldsymbol{\gamma}_{ki} - \sum_i \boldsymbol{\gamma}_{ik} \right] \leq 0 \tag{3d}$$

$$\delta_j^k \geq \lambda_{pj}^k, \ \forall p,k \tag{3e}$$

$$\sum_j \lambda_{pj}^k w_{pj} = \pi_{pk}, \ \forall p,k \tag{3f}$$

$$\sum_j \lambda_{pj}^k y_{pj} \geq A_k \left(v_{pk} + e_{pk} \pi_{pk} \right), \ \forall p,k \tag{3g}$$

$$\sum_j \lambda_{pj}^k = S_k, \ \forall k \tag{3h}$$

$$\sum_k \sum_i \left[\sum_p \sum_q \mathbf{c}_{ki}^{pq} \boldsymbol{\alpha}_{ki}^{pq} + \mathbf{d}_{ki} \boldsymbol{\gamma}_{ki} \right] + \sum_k \sum_p \mathbf{c}_X \boldsymbol{\mu}_{pk} + \sum_k \mathbf{c}_Z \boldsymbol{\eta}_k \leq C \tag{3i}$$

$$\sum_i \sum_q \boldsymbol{\alpha}_{ki}^{pq} \leq \mathbf{g}_k^p, \ \forall p,k \tag{3j}$$

$$\sum_k \sum_i \sum_p \sum_q \mathbf{I}_{ki}^{pq} \boldsymbol{\alpha}_{ki}^{pq} \leq \mathbf{u}_x \tag{3k}$$

$$\sum_i \boldsymbol{\gamma}_{ki} \leq \mathbf{f}_k, \ \forall k \tag{3l}$$

$$\sum_k \sum_i \mathbf{I}_{ki} \boldsymbol{\gamma}_{ki} \leq \mathbf{u}_z \tag{3m}$$

In constraint (3a) the decision variables $\boldsymbol{\alpha}_{ki}^{pq}$ represent the quantity of inputs transferred from process p of DMU k to process q of DMU i; similarly, $\boldsymbol{\alpha}_{ik}^{qp}$ represents the quantity of inputs transferred from process q of DMU i to process p of DMU k. The difference between these two quantities represents the outflow of inputs from process p of DMU k towards all other DMUs. Clearly, DMU k cannot transfer to other DMUs more inputs than its endowment $\Sigma_k \Sigma_p \mathbf{x}_{pk}$ and this justifies constraint (3b), which is a statement of this feasibility condition for the reallocation of inputs. The non-negative decision variables μ_{pk} represent additional inputs that can be provided to process p of DMU k by the system. The provision of these additional resources comes at a unit cost of $\mathbf{c}_X$ and is allocated optimally by the program to the different nodes of the production network.

One scenario we consider is the case in which the cost of this provision is prohibitive, de facto constraining these decision variables to be equal to zero. Constraints (3c) and (3d) have a similar interpretation, although they are expressed in terms of the inputs $\mathbf{Z}$ for which we do not observe allocation across the different services. Again, one scenario we consider is one in which the unit cost $\mathbf{c}_Z$ of providing additional inputs is prohibitive. Constraints (3e), (3f), (3g), (3h) have already been discussed in connection with program (1). Constraint (3i) takes into account the cost of reallocating inputs and the cost of providing new inputs to the system. C is the overall available "expenditure" the system can afford. The cost of reallocating inputs plus the cost of the provision of additional inputs cannot be higher than this given cost C. $\mathbf{c}_{ki}^{pq}$ is the row vector of unit costs of reallocating inputs $\mathbf{x}$ from process p of DMU k to process q of DMU i. $\mathbf{d}_{ki}$ is the row vector of unit costs for reallocating the inputs $\mathbf{z}$ from DMU k to DMU i. $\mathbf{c}_X$ is the row vector of unit costs for the provision of the additional inputs $X = \Sigma_k \Sigma_p \boldsymbol{\mu}_{pk}$ and $\mathbf{c}_Z$ is the row vector of unit costs for the provision of the additional inputs $\Sigma_k \boldsymbol{\eta}_k$. The quantities of additional inputs $(\boldsymbol{\mu}_{pk}, \boldsymbol{\eta}_k)$ are decision variables in the program. Note that costs do not need to be dollar costs. For example, a scenario we consider later is an increase in the quantity of inputs $\mathbf{z}$. To this purpose, we can set C as the overall number of additional units of inputs provided to the system and by setting the unit costs to one, we can model the provision of a given quantity of additional inputs. Note that by making the quantity of additional resources a decision variable we can model trade-offs between the reallocation of existing resources and the provision of additional resources as two alternative strategies to reach the same efficiency target.

We use constraints (3j), (3k), (3l) and (3m) to model potential non-linearities in the cost of reallocation. Constraints (3j) and (3l) provide an upper bound for the quantity of inputs that can be reallocated from DMU k to other DMUs. One case we consider is to allow only a reallocation of 5% or 10% of the inputs, to avoid disruptions to the ordinary operations of the court. This can be accomplished by setting, for instance, $\mathbf{f}_k = 0.05 \mathbf{z}_k$.

Table 1 Inputs and outputs (Avg. 2005–2012)

Variables	(Service)	Mean	St. Dev.	Min.	Max
Outputs (Y)					
Defined cases	(Civil)	16,358.4	24,722	1329.4	196,114.5
Defined cases	(Criminal)	7416.5	9026.3	455.3	64,437.9
Inputs (X)					
Non-professional Judges	(Civil)	3	4.2	0	29.1
Non-professional Judges	(Criminal)	0.7	1.4	0	7.3
Inputs (Z)					
Professional Judges	(Non-allocated)	30.6	48.1	6	379
Administrative Staff	(Non-allocated)	92	124.4	13.6	1083.4
Non-professional Judges	(Non-allocated)	0.4	0.8	0	4.5
Inputs (W)					
Pending cases	(Civil)	21,099.3	31,360.9	1363.4	216,083.6
Pending cases	(Criminal)	6889.1	9777.8	368	66,324.9
Others					
Population	–	361,813.3	398,311.2	2052	2,761,477

Reallocating above this level may involve a higher re-organizational cost than the one modeled through constraint (3i). Constraints (3k) and (3m) allow us to model costs associated with the relocation of personnel and potential re-training costs. One case we consider is to use these constraints to set forbidden routes for the reallocation of inputs, by only allowing the reallocation of inputs within a 50 km (case I) or a 100 km (case II) radius of the current location of inputs (personnel). $\mathbf{I}_{ki}^{pq}$ is the row vector of unit costs for the reallocation and $\mathbf{u}_x$ the maximum allowable cost. This program is solved under alternative policy scenarios as described in the results section (cases III and IV). These alternative scenarios involve specifying some values for the parameters of the program.[10]

3 Data

We consider courts of first instance (*Tribunale Ordinario*), which have jurisdiction over civil and criminal cases. Generally presided over by one judge, for important cases a panel of three judges presides. Their decisions can be appealed at the *Corte d'Appello* (for reasons of substance, i.e., concerning facts giving rise to the case) or at the *Corte di Cassazione* (i.e., for reasons concerning legitimacy or similar issues). We refer to a panel of 165 courts (the Italian court population) for the years 2005–2012.

The following measures were used for inputs and outputs: for outputs the total number of civil and criminal cases completed in a given year; for inputs, the personnel (professional and non-professional judges, administrative staff) and the number of pending civil and criminal cases at the beginning of the year. Using pending cases as an input was first suggested by Lewin et al. (1982), and can be defended on common sense grounds: without pending cases, there are no processed cases and therefore no output. In general, in any system, there is a percentage of pending cases, and these can be interpreted as an intermediate input stock (raw material inventory or working capital). We should point out that the number of pending cases is included in our model as a congestion factor: this means that given the number of judges, the number of processed cases increases if the number of pending cases increases, but there is a limit after which congestion may kick in and reduce the number of processed cases.

Table 1 provides descriptive statistics for the inputs and outputs available for the pooled sample of 165 courts over the period 2005–2012 (a total of 1320 observations). On average, an Italian court completes almost 24,000 cases per year, with quite a wide range between courts (the minimum is fewer than 2000 cases, the maximum is above 260,000 cases). On average a court has 31 judges, from 6 to almost 400. The stock of pending cases at the beginning of the year is around 21,000 civil and about 6900 criminal cases, from a minimum of fewer than 2000 overall cases to almost

[10] We have incorporated the impact of break-ups as a separate policy scenario given their importance in the Italian system. The reason for doing so is empirical not theoretical or methodological. In the Italian context, the effect of having a few very large courts of justice operating at a massively unproductive scale is quite substantial. As we pointed out elsewhere, this accounts for more than a third of the overall inefficiency of the system. Considering that this is connected to a simple policy of break-ups of large courts, its effect deserves to be studied separately. Just to give an example, merging two small units may require personnel relocating to a different city. Breaking up a large court of justice, for instance in Rome, does not require personnel to be moved intercity. Indeed it does not necessarily require new buildings, since judges can keep operating in the same building under a different administrative unit. Courts have become bloated well beyond the size that is technically efficient in terms of scale essentially for political reasons.

280,000 cases over the period considered. As a very rough measure, because we consider the average for all the courts over the years, disposition time is about 15.5 months for civil and 11.1 for criminal cases.[11]

4 Empirical results

We now set out the results of the computations of the various programs, by discussing the counterfactual analysis of the different policy scenarios. Further results, including a breakdown of courts and their distribution across geographical areas, are provided in the online Appendices.

4.1 Counterfactual analysis of supply policies

Four sets of policies—a) the introduction of best practices, b) the break-up of large courts, c) the reallocation of personnel (judges and administrative staff), and d) an increase in personnel—plus their combinations, can be considered tools on the supply side, the effects of which can be computed with our proposed models. Some of these policies have now been implemented (after the period examined here).

Apart from the Pinto law of 2001 establishing damages for lengthy cases, most of the measures until 2012 were designed to reduce case inflow. They include an increase in court fees to avoid excessive recourse to courts, tougher criteria for appealing, the introduction of alternative dispute resolution (ADR) mechanisms, and so on (Esposito et al. 2014). Further changes have been proposed, such as backlog-reducing teams and other measures previously introduced in pilot schemes[12] to be extended to courts throughout Italy. But overall the need for better court management to handle cases more actively, with data systems and performance accountability, is recognized (Esposito et al. 2014: 13).

One example of a good *case management practice*, often cited, is the Strasbourg method adopted first by the Turin court and subsequently extended elsewhere (Caponi, 2016). This method is based on the active leadership role of the President of the court, making each judge responsible for reaching clear and transparent objectives, to be monitored actively, and changing case management from 'last in - first out' (LIFO) to "first in–first out" (FIFO) (Abravanel et al. 2015). The *Consiglio Superiore della Magistratura* (CSM), the self-governing judicial body, has recently started to introduce best practices such as strategy and planning using a formal "organizational document", working trials in sequence and not in parallel and the intensive introduction of IT technologies.[13] Effective practices include case-flow management and the production of statistics, areas in which Italy has been lagging behind compared to other OECD countries (see, e.g., Palumbo et al. 2013: 34–35), certainly in the years under consideration.

Another possible policy is the re-design of *court geography* (Bartolomeo, 2013). The geography of Italian courts was originally designed after Italian unification in 1865, and underwent a number of changes during the fascist period, after World War II and in the late nineties. In the early nineties, the CSM suggested the need to break up large courts such as Rome, Naples, Milan and Turin (CSM, 2010).[14] However, most of the literature on Italian court efficiency has highlighted increasing returns to scale and thus the need to merge courts,[15] leading to similar policy suggestions offered by the Ministry of Finance[16] and eventually implemented by the Monti Government by reducing the number of courts by about 20%.[17]

Last, an increase in judges and administrative staff is another policy option, but given the poor state of public finances in Italy and over-staffing compared to other OECD countries (Palumbo et al. 2013), it has been difficult to implement. This policy, however, together with the break-up of large courts, makes a total of 32 possible policy scenarios. For the sake of clarity and available space, we first consider the scenarios with the given personnel. [18] For each scenario, illustrated in Table 2,

[11] Disposition time (in years) is calculated as $\frac{pending\ cases}{completed\ cases} = \frac{21{,}099}{16{,}358} = 1.27$, or 15.5 months for civil cases.

[12] "Since 2004, the EU supported a roll-out of the Turin and Bozen courts' experience to the entire country (Program Title: Diffusion of best practices in the Italian Judicial Offices). This program made some progress (e.g., for the Milan Court). However, the program faced implementation constraints as well as jurisdictional issues between regional and central authorities. The central government has taken a stronger role in program management since 2010–2011, with the Ministry of Public Administration setting up an effective central monitoring system in 2011 and the Ministry of Justice putting in place professional management in 2012. This helped secure the EU structural funds" Esposito et al. 2014: 10).

[13] For a more detailed explanation see, e.g., www.csm.it/web/csm-internet/il-progetto-buone-prassi/il-fenomeno-buone-prassi.

[14] The CSM "suggested a split of their structures on a territorial basis, dividing their district into two or three parts with corresponding court and district attorney for each of them" (CSM, 2010: 4).

[15] A notable exception is represented by Peyrache and Zago (2016): using Italian court data for the period 2003–08, they find that the breaking up of large courts could reduce aggregate inefficiency by 22%.

[16] The Ministry of Finance estimated the elasticity of scale of Italian courts using 2006 data, finding that about 85% of courts were too small and confirming earlier results for 1996 and 2001 set out by Marchesi (2003, 2008). Therefore, the policy recommendation was "…to revise judiciary geography, by merging the smaller courts in order to realize economies of scale and specialization…"(CTFP 2008: 46).

[17] With Legislative Decree 155 dated 7 September 2012, the Monti government merged 26 small courts (out of 165 at the national level) into larger, adjacent courts, taking effect in 2013.

[18] More recently, to benefit from the EU public funds associated with the Recovery Plan following the Covid pandemic, the Italian Government designed a reform of the justice system which, among other things, increase the number of judges and other personnel to be employed in the courts across Italy.

Table 2 Possible supply policy scenarios

	No Break-ups		Break-ups	
	Current efficiency (i)	Full efficiency (ii)	Current efficiency (iii)	Full efficiency (iv)
No Reallocation	1	2	3	4
Constrained Reallocation (I) (5%—50 km)	5	6	7	8
Constrained Reallocation (II) (10%—100 km)	9	10	11	12
Full Reallocation	13	14	15	16

we show the overall average equilibrium disposition time, while its distribution (across courts and geographical areas) is presented in the online Appendices.

4.1.1 With existing personnel

Table 3 shows the current average disposition time for the system as a whole (calculated as the total number of pending cases in the system over the total number of completed cases, i.e., the weighted average of the observed completion times of individual courts) and the average disposition time associated with different policy scenarios. Note that the *current* disposition time for the system is 15.5 and 11.1 months for civil and criminal cases, while the *equilibrium* disposition time—after the system adjusts to the steady state (see program of Eq. (2))—is 14.6 (civil) and 7.3 months (criminal cases). These figures are a weighted average for the whole system.

Table 3 also illustrates the equilibrium outcome of the different policy scenarios.[19] Overall, the single most effective policy would be the introduction of best practices (scenario 2), leading to a reduction of disposition time to 11.9 and 6.3 months (respectively for civil and criminal cases). This is followed by the reallocation of personnel (scenario 13, with a reduction to 12.3 and 6.7 months, respectively). Next, changing the geography of courts by breaking up larger ones (scenario 3) would lead to a reduction of trial length to 13 and 5.9 months. Last, constrained reallocation I and II (scenarios 5 and 9) would give results similar to break-ups, leading to 13.5–13.2 and 6.8–6.6 months, respectively.

Having the chance to adopt two policy tools, the best combination would be the use of best practices, either together with break-ups (scenario 4, leading to a further reduction to 10.7 and 5.2 months), or the reallocation of personnel (scenario 14, leading to an average disposition time of 9.6 and 5.6 months). Only by implementing these three policies together—presumably a rather challenging task—the system would be taken to steady state disposition times comparable to other OECD countries (8 and 4.7 months; scenario 16).

4.1.2 Increasing available personnel

Consider now the effects of an increase in personnel (professional and non-professional judges and administrative staff), as such and in combination with other policies. Table 4 shows the average disposition times of different policy combinations. First, increasing the available personnel as a stand-alone policy would have a substantial impact on reducing disposition times: by increasing personnel by 10%, for instance, the average disposition time in the system would fall to about 13 and 6.4 months (for civil and criminal cases; Table 4) from the equilibrium status quo of about 14.6 and 7.3 months (as seen in Table 3). Note that this impact is similar to a redefinition of court geography, i.e., break-ups of large courts (scenario 3 seen before).

An interesting comparison can be made between the increase of personnel as such and in combination with other policies, e.g., together with its optimal reallocation. The most impactful combination would be with best practices (reduction to 10.5 and 5.8 months respectively for civil and criminal cases), then with full reallocation (10.7 and 6.3 months), break-ups of bigger courts (11 and 5.5 months), and constrained reallocation I and II. Last, all policies combined would lead to 7.2 and 4.9 months of trial duration for civil and criminal cases.

We also consider an increase of personnel by 25% taking into account the relative cost of each type of personnel.[20] By increasing personnel by 25%, together with a constrained reallocation (case III) would lead to a fall of disposition times to 10.3 and 6.4 months respectively, and to 10.2 and 6.4 months in case IV, i.e., with a reallocation for each court of at most 10% of personnel within 100 km. Note however that increasing personnel from 10 to 25% would have no effect were the other three policies already adopted, i.e., the average disposition time would remain at 7.2 (civil) and 4.9 (criminal) months.

Next, we consider how the increased personnel should be optimally split into different categories (professional and

[19] For example, a policy of breaking up large courts together with an optimal reallocation of personnel (policy scenario 15) would reduce the disposition time of the system from 14.6 to 9.9 months for civil cases (and from 7.3 to 5.2 months for criminal cases). This would result from the optimal use of scale economies and the unconstrained reallocation of personnel.

[20] We assume a cost equal to 1 for an additional judge, equal to 0.5 for an additional administrative member of staff, and equal to 0.1 for the reallocation of a judge (see the constraints of Eqs. (3k) and (3m)).

Table 3 Average disposition time (in months) for policy scenarios

	Cases	No Break-ups		Break-ups	
		Current efficiency (i)	Full efficiency (ii)	Current efficiency (iii)	Full efficiency (iv)
No Reallocation	Civil	14.6	11.9	13.0	10.7
	Criminal	7.3	6.3	5.9	5.2
Constr. Realloc. (I)	Civil	13.5	11.1	11.8	9.8
(5% - 50 km)	Criminal	6.8	5.9	5.6	4.9
Constr. Realloc. (II)	Civil	13.2	10.8	11.2	9.4
(10% - 100 km)	Criminal	6.6	5.8	5.5	4.8
Full Reallocation	Civil	12.3	9.6	9.9	8.0
	Criminal	6.7	5.6	5.2	4.7

Observed disposition time: civil 15.5, criminal 11.1 months.

Table 4 Increasing Personnel ($+10-25\%$) - Average disposition time (in months)

	Cases	No Break-ups		Break-ups	
		Current efficiency (i)	Full efficiency (ii)	Current efficiency (iii)	Full efficiency (iv)
		$\boldsymbol{\Delta = +10\%}$			
No Reallocation	Civil	13.0	10.5	11.0	9.1
	Criminal	6.4	5.8	5.5	4.9
Constr. Realloc. (I)	Civil	11.6	9.5	9.7	8.1
(5% - 50 km)	Criminal	6.3	5.6	5.5	4.8
Constr. Realloc. (II)	Civil	11.3	9.3	9.4	7.8
(10% - 100 km)	Criminal	6.3	5.6	5.5	4.8
Full Reallocation	Civil	10.7	8.6	8.8	7.2
	Criminal	6.3	5.8	5.4	4.9
		$\boldsymbol{\Delta = +25\%}$	& **Trade-off**		
Constr. Realloc. (III)	Civil	10.3	8.5	8.7	7.3
(5% - 50 km)	Criminal	6.4	5.7	5.6	4.9
Constr. Realloc. (IV)	Civil	10.2	8.4	8.6	7.2
(10% - 100 km)	Criminal	6.4	5.8	5.6	4.9

Observed disposition time: civil 15.5, criminal 11.1 months.

non-professional judges and administrative staff) in different scenarios (see Table 5). Using only the "increase personnel by 10%" policy means adding about 565 people, including 396 professional judges, 82 administrative personnel and remaining 87 non-professional judges in the optimal allocation. When in combination with other policies, professional judges may be increased less: by 337 when best practices are introduced, and by 321 with break-ups. When in combination with reallocation, increasing personnel would translate more substantially (or completely, see column I) into more professional judges and less or no additional administrative staff. Similarly, with an increase of 25% in personnel: in brief, the more policy tools are implemented, the fewer the additional professional judges required.

5 Concluding remarks

Justice delayed is justice denied. The time needed to provide a service is an important benchmark in many settings. However, ensuring fast delivery times is costly, thus taking into account the resources needed for fast provision is paramount. Using a rather general optimization model to quantify the trade-off between resource utilization and service time for Decision-Making Units operating multiple services, we extend the standard practice of considering disposition times and carry out an efficiency analysis investigating the sources of inefficiency of Italian courts of justice, where trials have historically been very long. After taking into consideration the feedback effect from faster disposition times and considering the potentially congesting

Table 5 Increasing Personnel (+10 − 25%) - Additional personnel (in units)

		No Break-ups		Break-ups	
	Personnel	Current efficiency (i)	Full efficiency (ii)	Current efficiency (iii)	Full efficiency (iv)
		Δ = + 10%			
No Reallocation	Judges	396	337	321	291
	Administrative	82	131	119	140
Constr. Realloc. (I)	Judges	453	406	375	336
(5% - 50 km)	Administrative	22	62	68	101
Constr. Realloc. (II)	Judges	474	447	396	352
(10% - 100 km)	Administrative	0	21	46	84
Full Reallocation	Judges	505	505	494	483
	Administrative	0	0	0	0
		Δ = + 25%	**& Trade-off**		
Constr. Realloc. (III)	Judges	848	830	782	771
(5% - 50 km)	Administrative	37	44	71	78
Constr. Realloc. (IV)	Judges	884	844	813	785
(10% - 100 km)	Administrative	590	667	668	710

Observed disposition time: civil 15.5, criminal 11.1 months.

effect of large stocks of pending cases, we investigate the impact of different reform policies of the court system on the average disposition times for the overall system and on its break-down in different districts.

We find that the single most effective policy would be the adoption of best practices by the courts, which could reduce the average overall time to complete a trial by about one quarter, from 15.5 to 11.9 months for civil (and from 11.1 to 6.3 months for criminal) cases. An alternative policy would be the combination of court geography redesign together with the reallocation of the existing personnel according to best use, leading to a reduction to about 9.9 months (5.2 for criminal cases). Finally, with existing judges, combining the adoption of best practices with the break-up of courts and optimal reallocation of personnel, the average disposition time of the system would be almost halved (at 8 and 4.7 months), even accounting for the increased demand for justice resulting from faster disposition times.

While the costs of these supply policies, taken individually or in combination, may be difficult to ascertain, the cost of increasing personnel is easier to calculate. Thus, the alternative policy of increasing the number of judges by about 25% (and their optimal allocation) would have comparable effects to the implementation of best practices in combination with the resizing of the courts or with a constrained reallocation of personnel, with a total initial cost of at least 100 million euros per year. These alternative policy scenarios would be sufficient to bring the system down to a disposition time comparable to other OECD countries, and the benefits of these policies would be substantial, as court inefficiencies account for a loss of about 1% in GDP (Draghi 2011).

The paper does not consider two further issues that could be the basis for future research. The first is an alternative scenario in which the number of incoming cases is reduced by introducing some sort of out-of-court settlement process. This has occurred in recent years, for instance through alternative dispute resolution (ADR) mechanisms introduced to reduce the use of the courts. The second issue not addressed here is the transition towards the steady-state. For example, when the system converges from the observed disposition time of 15.5 months (for civil cases) to a steady-state disposition time of 14.6 months, we do not specify the timing of this adjustment. Future research could explore these transition dynamics in more depth by designing appropriate transition scenarios.

Acknowledgements We would like to thank J. Bos, G. D'Inverno, U. Nizza, M.C. Silva, and participants at EWEPA2017 (London), NAPW2018 (Miami) and COST2022 (Porto) for useful inputs. The usual disclaimer applies. This paper was written while Zago was visiting the School of Economics at the University of Queensland, whose hospitality is gratefully acknowledged. The research (by Zago) leading to these results received funding from PIA at the University of Verona.

Funding Open access funding provided by Università degli Studi di Verona within the CRUI-CARE Agreement.

Compliance with ethical standards

Conflict of interest The authors declare no competing interests.

6 Appendix A: The Demand Feedback effect

The first step is to estimate the demand for justice. For this purpose, the following regression model is used:

$$\log inc_{jt} = \alpha_j + \mathbf{X}_{jt}\boldsymbol{\beta} - \gamma \log t_{jt-1}, \quad (4)$$

where $j = 1, \ldots, J$ stands for the courts and $t = 1, \ldots, T$ is the reference year. The variable inc_{jt} is the number of incoming cases in year t and court j, and $\mathbf{X}_{jt}$ contains control variables, including individual level dummy variables and the population size of each court district (this is exogenously determined, since each district population refers to a given court and case movements between courts is forbidden by law). The average disposition time for a new case is given by the ratio of the number of pending cases to the number of completed cases $t_{jt} = w_{jt}/y_{jt}$.[21]

The demand equation includes the lagged value for the time to complete trials. This can be interpreted both as a causal relationship between completion time and the number of incoming cases (quantity demanded) on the assumption of adaptive expectations for the plaintiff, or as a prediction equation for the number of incoming cases. To check the robustness of our estimates, we also tested the sign and size of demand elasticity (γ) by using the lagged value of the average disposition time as an instrument for the following simultaneous demand equation (based on rational expectations):

$$\log inc_{jt} = \alpha_j + \mathbf{X}_{jt}\boldsymbol{\beta} - \gamma \log t_{jt}. \quad (5)$$

The steady state equilibrium condition specifies that the number of processed cases each year must be equal to the number of incoming cases:

$$inc_{jt} = y_{jt}. \quad (6)$$

[21] Since we also have data for the number of judges (g_{jt}) and the number of completed (y_{jt}) and pending cases (w_{jt}), we can compute the average queue disposition time for each court in each time period.

This equilibrium condition, together with the material balance condition (described above) means that in equilibrium the queue for the court is in steady state with the number of pending cases constant from one year to the next, which in turn (unless some of the control variables have changed from one year to the next) means that $inc_{jt} = inc_{jt-1}$.

The equilibrium condition and material balance condition mean that we can derive a relationship between the number of pending cases and the number of completed cases from the demand function:

$$y = w^{-\frac{\gamma}{1-\gamma}} \exp(\mathbf{X}\boldsymbol{\beta}/(1-\gamma)).$$

This equation is useful in order to analyze demand trade-offs in the pending-completed cases space. This also implies the following equilibrium relationship between the average disposition time and the number of pending cases (dividing the previous equation by p and taking the inverse):

$$t = \frac{w^{\frac{1}{1-\gamma}}}{\exp(\mathbf{X}\boldsymbol{\beta}/(1-\gamma))},$$

where $\mathbf{X}\boldsymbol{\beta}$ is the prediction based on the demand regression estimates.

Unless demand is rigid in terms of disposition time ($\gamma = 0$), increasing the efficiency of production (by increasing the number of processed cases) decreases disposition time and increases the number of incoming cases. Therefore efficiency gains may be overestimated if the demand side is ignored.[22]

6.1 Demand estimation

We estimated a number of different specifications of the demand equation and show the results in Table 6 (civil) and 7 (criminal cases). For the demand equation based on *adaptive expectations*, we use *lagged* disposition time to assess the elasticity of demand, and population as a proxy for the size of demand. We consider OLS, Fixed Effects, Random Effects and Fixed Effects with time dummy variables. As can be inferred from the tables, both the Random Effects and OLS models are rejected in this specification. The estimated elasticity of demand is -0.495 (-0.182 for criminal cases) for the Fixed Effects model and -0.469 (-0.221 for criminal cases) for the Fixed Effects model that includes time dummy variables.

[22] In the estimations, as will be shown below, we find demand to be quite inelastic at a value of $\gamma = 0.495$ for civil and $\gamma = 0.182$ for criminal cases. This means that in our dataset a 10% increase in disposition time reduces the number of incoming cases by 4.95% and 1.8% respectively.

Table 6 Demand estimation results—civil cases

	Adaptive expectations				Rational expectations			
	1 (FE)	2 (FE1)	3 (OLS)	4 (RE)	5 (2SLS)	6 (Court FE)	7 (Year FE)	8 (Both FE)
Population	1.24[a]	1.245[a]	1.065[a]	1.038[a]	1.066[a]	0.552	1.066[a]	1.58[b]
	(0.23)	(0.292)	(0.010)	(0.025)	(0.009)	(0.431)	(0.009)	(0.517)
Lagged compl. time	−0.495[a]	−0.469[a]	0.323[a]	−0.297[a]	IV	IV	IV	IV
	(0.025)	(0.026)	(0.019)	(0.025)				
Current compl. time	–	–	–	–	0.336[a]	−1.107[a]	0.342[a]	−1.0546[a]
					(0.019)	(0.104)	(0.019)	(0.103)
Constant	–	–	−1.776[a]	−1.582[a]	−1.78[a]	–	–	–
			(0.052)	(0.134)	(0.05)			
Year FE	–	Yes	–	–	–	–	Yes	Yes
Court FE	Yes	Yes	–	–	–	Yes	–	Yes
Adj. R-squared	0.999	0.999	0.916	0.63	0.921	0.997	0.999	0.997
No. observations	990	990	990	990	990	990	990	990

[a] = 1% s.l.

[b] = 5% s.l.

Table 7 Demand estimation results—criminal cases

	Adaptive expectations				Rational expectations			
	1 (FE)	2 (FE1)	3 (OLS)	4 (RE)	5 (2SLS)	6 (Court FE)	7 (Year FE)	8 (Both FE)
Population	0.394	0.39	0.961[a]	0.963[a]	0.961[a]	−0.020	0.961[a]	0.304
	(0.465)	(0.564)	(0.011)	(0.025)	(0.011)	(0.56)	(0.011)	(0.729)
Lagged compl. time	−0.182[a]	−0.221[a]	0.008	−0.125[a]	IV	IV	IV	IV
	(0.022)	(0.022)	(0.016)	(0.019)				
Current compl. time	–	–	–	–	0.009	−0.324[a]	0.003	−0.411[a]
					(0.017)	(0.048)	(0.017)	(0.052)
Constant	–	–	−1.487[a]	−1.514[a]	−1.487[a]	–	–	–
			(0.062)	(0.137)	(0.062)			
Year FE	–	Yes	–	–	–	–	Yes	Yes
Court FE	Yes	Yes	–	–	–	Yes	–	Yes
Adj. R-squared	0.995	0.999	0.86	0.56	0.861	0.999	0.999	0.999
No. observations	990	990	990	990	990	990	990	990

[a] = 1% s.l.

[b] = 5% s.l.

We consider the Fixed Effects model to be the best specification in this context since it enables us to control for unobserved heterogeneity and at the same time is more parsimonious than the model including time dummies. Both the OLS and Random Effects models are excluded because they produce inconsistent estimates of demand elasticity.

The last four models in Tables 6 and 7 consider the specification under *rational expectations* where the quantity demanded is a function of the *current* disposition time,

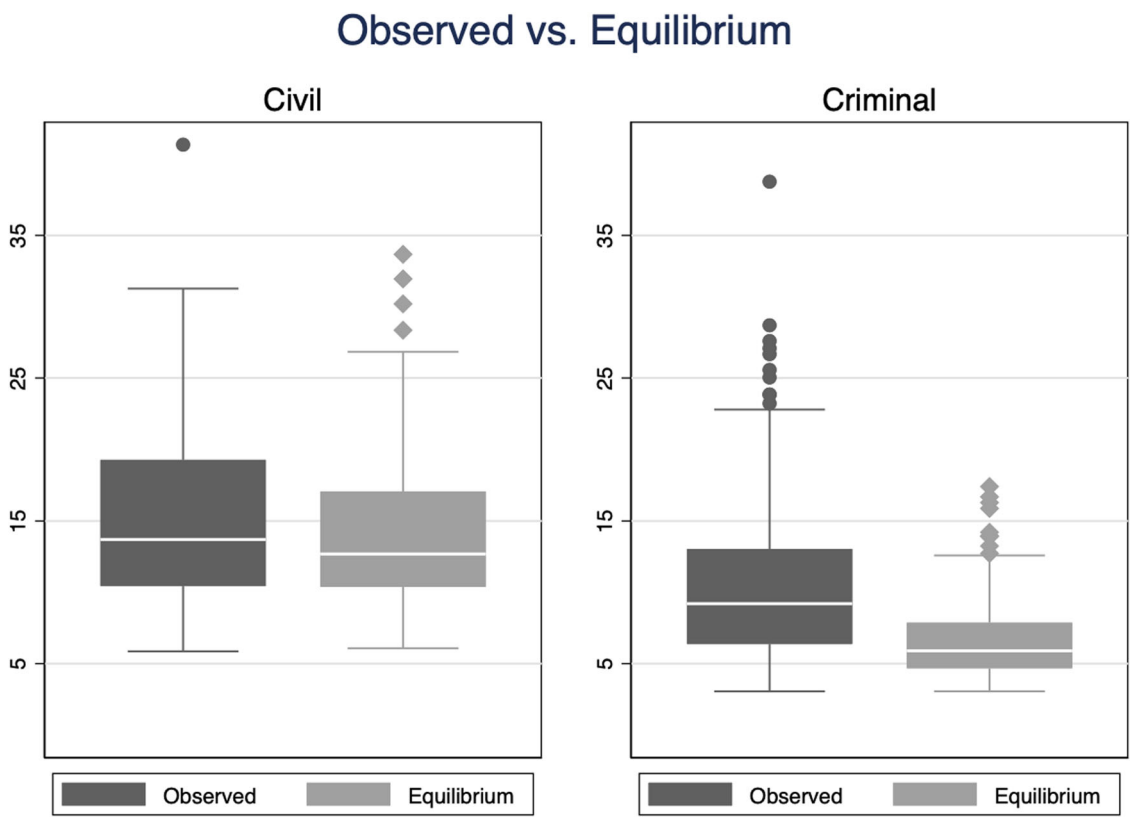

Fig. 2 Observed vs. equilibrium disposition time

rather than the lagged disposition time. We estimated such models by using the lagged disposition time as an instrument for the current disposition time. The 2SLS estimator is shown as well as the same estimator using court and time dummies. As for the previous set of regression models, the models that do not include court effects (i.e., without unobserved heterogeneity) produce inconsistent estimates. Models 6 and 8 include court fixed effects and the associated estimated elasticity is a consistent estimator of the population quantity. The estimated elasticity for model (6) is −1.107 (−0.324 for criminal cases), while the same elasticity in model (8), which includes time dummies as well, is −1.055 (−0.411 for criminal cases). These estimates possess a larger elasticity of demand compared to the adaptive expectations estimations. In terms of the trade-offs between parsimony and data fitting, we consider the fixed effects model with adaptive expectation to be the best of the 8 estimated models. Therefore we use an elasticity of demand of −0.495 (civil) and −0.182 (criminal cases) for the subsequent analysis.

Based on these predictive models, we linearize the demand feedback effect by taking a first-order Taylor expansion at the observed level of pending cases. Although the coefficients of the predictive model are common to all courts of justice, the coefficients of the Taylor expansion will be different for different courts. We store the intercept v_{pk} and the slope e_{pk} coefficients for each court of justice and each service process (civil and criminal) as additional data to be used in our optimization model. Since the elasticities of demand are quite rigid, the error in the prediction of the demand will be second-order compared to the gains in terms of computation complexity of the optimization programs.[23]

7 Appendix B: Further results with existing personnel

7.1 Impact across courts

Table 3 shows the current average disposition time for the system as a whole and the average disposition time associated with different policy reform scenarios, as explained in the main text of the paper. Now we look also at disposition times across courts, as shown in Fig. 2. In both

[23] There are other ways of linearizing the programs making use of a conservative (inflated) estimate of the demand feedback effect. Alternatively one may try to solve the associated convex program directly.

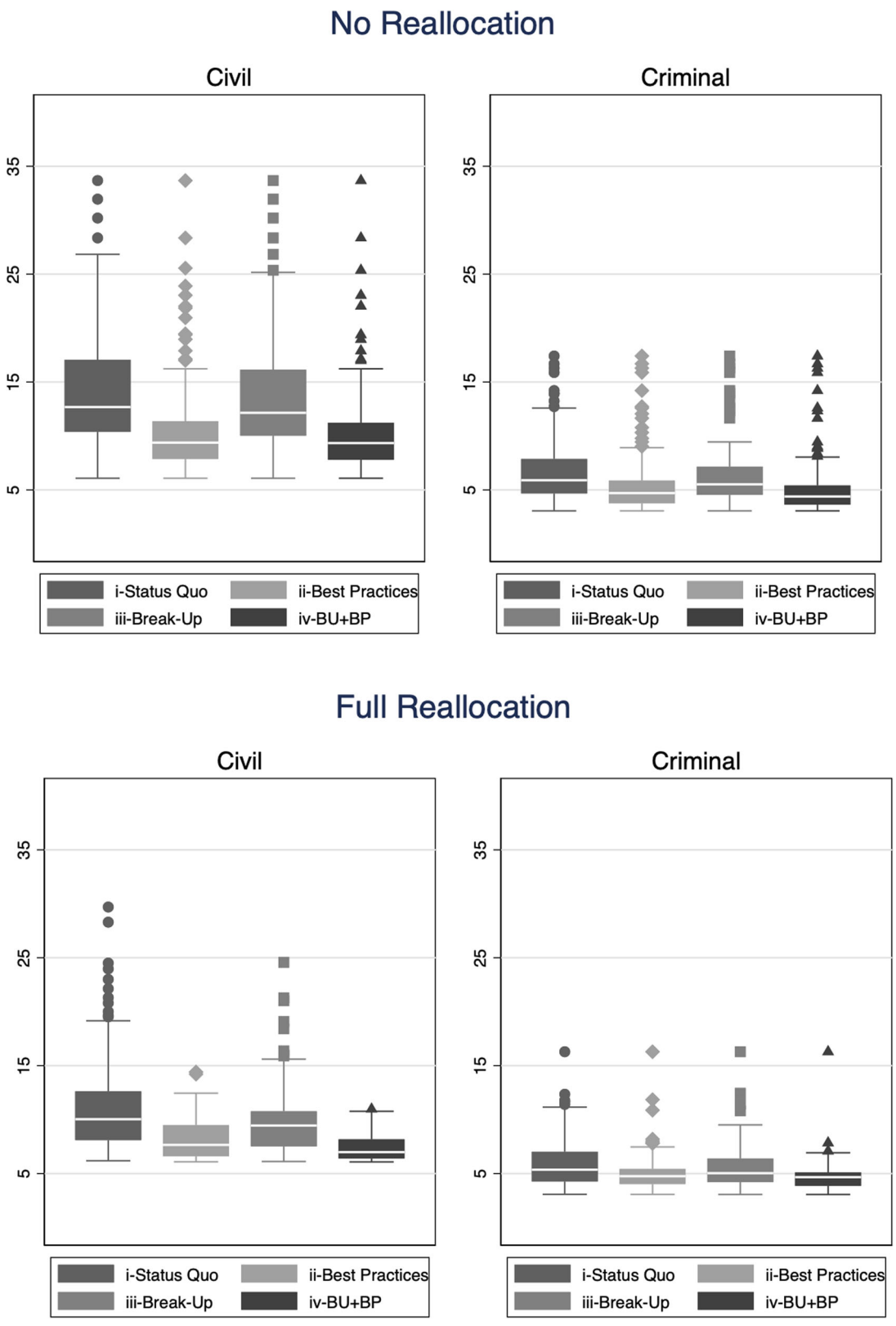

Fig. 3 Impact of policy scenarios (i) to (iv) with given personnel

panels, we see that moving from the observed to the equilibrium disposition times would reduce both their median and their dispersion, i.e., the equilibrium disposition time for some of the slowest processing courts diminishes as the system moves towards equilibrium. This seems more pronounced for criminal cases (right-hand panel).

The box plots in Fig. 3 are useful to investigate the distribution of disposition times after the implementation of different possible policies. We consider the civil cases first (left-hand panels). Without the reallocation of personnel (scenarios 1 to 4 of Table 2), the median disposition time is reduced, but also that of the slowest courts, to about 16 months (without considering the outliers) with best practices alone (scenario 2) or together with break-up of courts (scenario 4). Break-up alone, on the other hand, would not significantly reduce either the median or the maximum disposition times (about 25 months).

When personnel is reallocated without constraints (full reallocation, or scenarios 13 to 16 of Table 2) the median is decreased by a couple of months, and the slowest disposition times (again without considering the outliers) are decreased to about 18 months. If the optimal personnel reallocation is combined with the implementation of best practices, the median is further reduced to about 8 months and the slowest processing court to about 13 months. The break-up of bigger courts, together with full reallocation, on the other hand, would reduce the median and the dispersion compared to solely personnel reallocation. Last, note that the three policies combined would reduce the median disposition time to about 7 months and the slowest processing court would process civil cases in about 11 months, a result comparable to other developed countries.

The impact for criminal cases is similar, to the extent that the most effective policy is the adoption of best practices, but starting from faster disposition times. Without personnel reallocation, indeed, disposition times would fall from a median of 6 to about 4 months (best practices) or about 5 months (break-up). With personnel reallocation only, median disposition times would fall to 5 months; further introducing other policies would have limited impact on the median time to process a case, but would reduce the time needed for the slowest court to 7 or 8 months respectively with the introduction of best practices and break-ups. Last, looking at the outliers as well, with personnel reallocation the slowest court would take about 16 months to process a case, with or without other policies. So 16 months seems an unchangeable threshold, even though it is due to only a court.

To summarize, from Table 3 it is clear that the proper implementation of best practices has a major effect in reducing the overall average disposition times in the system. The combination with other policies would add to this; in particular, the implementation of the three policy tools, that is the introduction of best practices, the reallocation of personnel, and the break-up of the large courts, i.e., scenario 16, would bring the system disposition time to about 7 (civil) and 5 (criminal cases) months and would drastically reduce the dispersion in the system, with only one court of justice taking more than 10 months (15 months for criminal cases) to process a case. Implementing these three policies together may be challenging, but the impact would be rather substantial, taking the system to steady state disposition times comparable to other developed countries.

7.1.1 The geographical distribution of policy effects

The maps below represent the reduction in disposition times in courts in the different policy scenarios of Table 2, i.e., with existing personnel.[24] Figures 4 and 5 illustrate the effects of introducing best practices and the break-up of courts in terms of *reducing* the time needed to complete a case compared to the equilibrium disposition times, respectively for civil and criminal cases. First, note that in both the equilibrium (and actual) disposition times are in general longer in the South and in the Islands (Sicily and Sardinia; scenario (i), panels a, in both figures).

A comparison with panel b shows that the introduction of best practices (scenario 2, Table 2) would have a major impact in the same regions and possibly on courts where disposition times are longer, i.e., in Southern Italy, and in particular for civil cases, for which on average Italian courts are slower. In contrast, introducing a break-up of courts would have a limited effect when taken in isolation (scenario 3, panels c), or when added to the adoption of best practices (scenario 4, panels d). However, it appears that best practices and break-ups have an impact on different sets of courts, showing some complementary effects when looking at their geographical distribution. A similar picture emerges when considering the geographical impact of these policies together with the optimal reallocation of personnel.

To summarize, in terms of geographical distribution, introducing best practices would have the most substantial impact, especially in courts located in the South and in the Islands of Italy, which are overall more inefficient and slow in processing cases. As an alternative, or in addition, the reallocation of judges may be a second best, working more effectively when combined with best practices.

[24] Similar results (available upon request) are obtained also considering an increase in personnel.

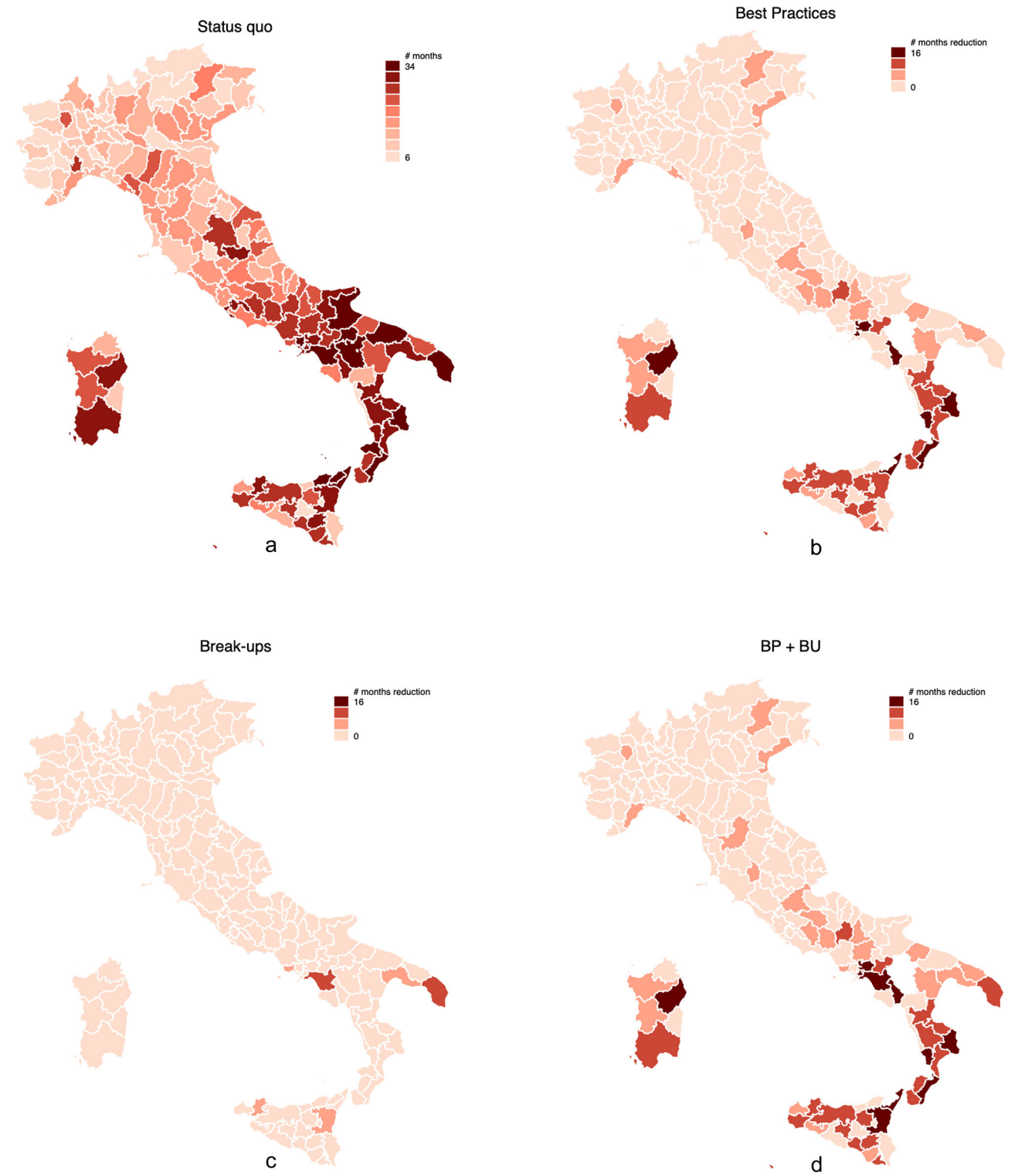

Fig. 4 No personnel reallocation—civil cases. **a** Scenario (i). **b** Scenario (ii). **c** Scenario (iv). **d** Scenario (iv)

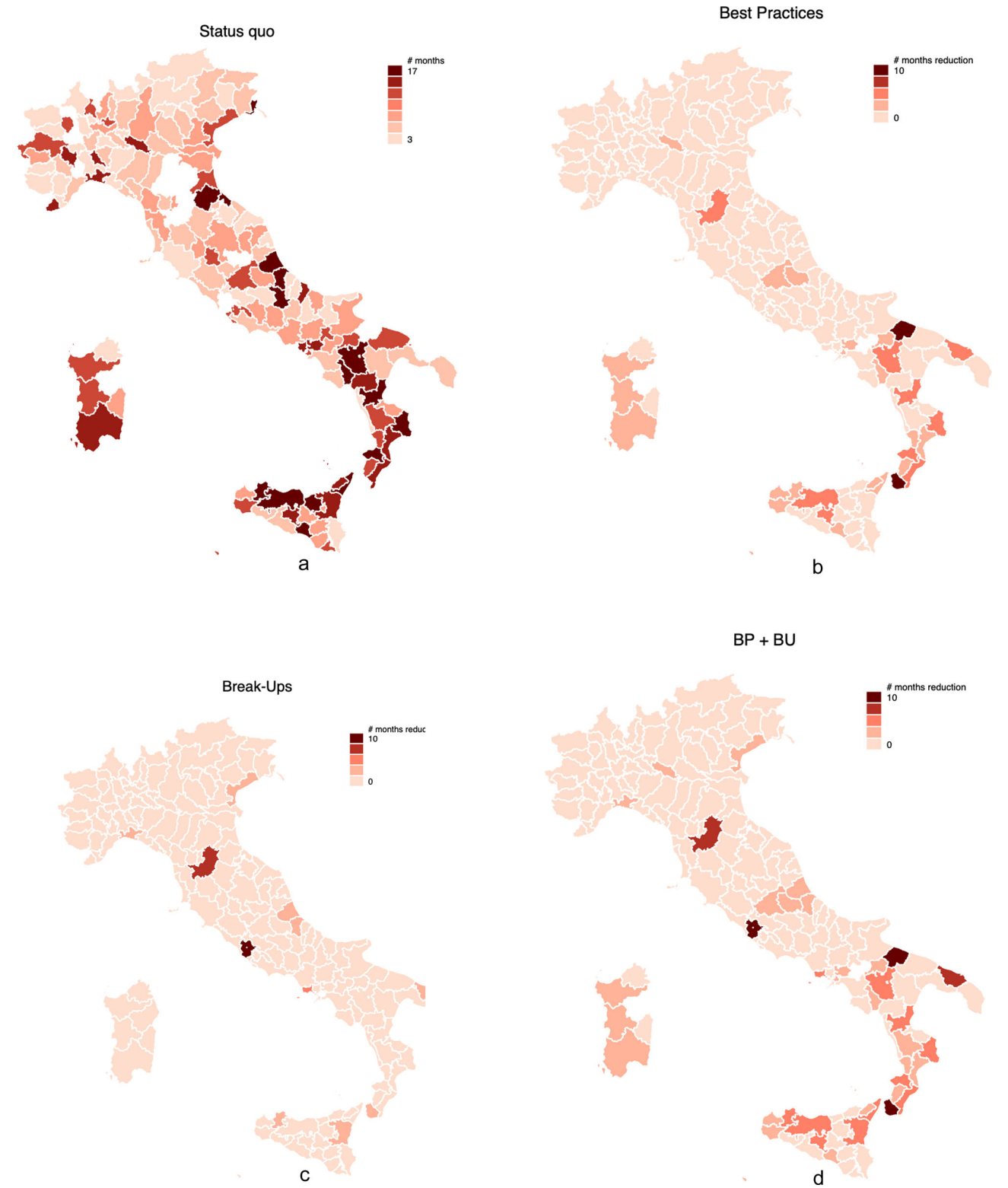

Fig. 5 No personnel reallocation—criminal cases. **a** Scenario (i). **b** Scenario (ii). **c** Scenario (iv). **d** Scenario (iv)

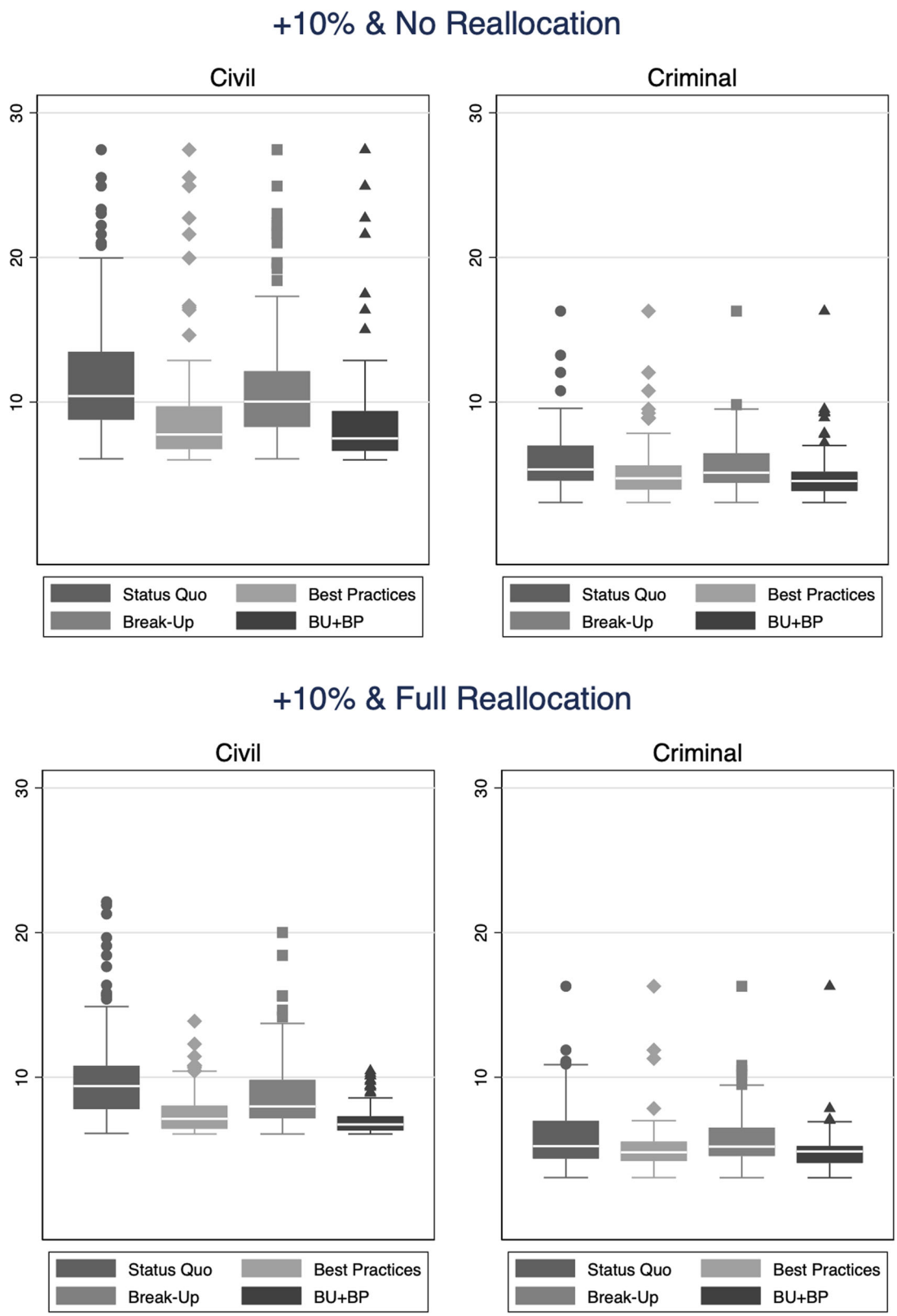

Fig. 6 Impact of policy scenarios with a 10% personnel increase

8 Appendix C: Increasing personnel

We now consider the effects of an increase in personnel, in combination with other policy scenarios. Table 4 shows the average disposition times—for civil and criminal cases—for different policy combinations, together with an increase of 10% and 25% in personnel.

8.1 Adding personnel to the current court system

First, we consider scenario (i), i.e., the increase in personnel without break-ups or the adoption of best practices. Considering civil cases first, note that by increasing personnel by 10%, the median disposition time in the system would fall to about 10.5 months (top-left panel of Fig. 6) from the equilibrium status quo of about 13 months (top-left panel of Fig. 3); moreover, the slowest court (excluding outliers) would reduce its disposition time from about 27 months to 20 months. An interesting pair of policies is the combination of increased personnel with its optimum reallocation (existing and newly hired; see the bottom-left panel of Fig. 6): it would reduce the median disposition time to about 9 months, but most importantly it would further reduce the time taken by the slowest court to 15 months.

Moving to criminal cases, by increasing personnel as a stand-alone policy the median disposition time would not decrease much, but the effect on the slowest court would be substantial, falling from about 13 months to less than 10 months (top-right panel of Fig. 6). Adding an unconstrained reallocation of existing and new personnel, however, would slightly raise the disposition times of the slowest courts—from less than 10 months to about 11 months—probably because it would work to the benefit of civil cases, "subtracting" production factors (judges and administrative staff) from criminal to civil cases (bottom-right panel of Fig. 6).

8.2 Together with the adoption of best practices

A relatively similar picture emerges when we consider the second policy combination, i.e., a personnel increase of 10% combined with the introduction of best practices (scenario (ii)). For civil cases, median disposition times would be reduced to about 8 months, with the slowest court taking about 13 months (but with outliers ranging at about 28 months). Further adding the optimal reallocation of personnel would reduce the median disposition time to about 7 months, bringing the slowest court to about 11 months (a more substantial reduction would be achieved for outlying courts, which would halve their disposition time from about 28 months to 14 months (see the bottom-left panel of Fig. 6).

For criminal cases, the effects would be more nuanced. Adding personnel (an increase of 10%) to best practices would not reduce median disposition times (in any case around 5 months), but the slowest court times would be reduced from 9 to about 8 months. A further reduction for the slowest court would principally result from implementing the additional policy of optimal personnel reallocation, with its disposition time reduced to about 7 months (bottom-right panel of Fig. 6).

8.3 Together with a new geography of courts

A further policy combination involves an increase in personnel combined with the optimal break-up of large courts (scenarios (iii)). First of all, the effect on civil cases is again more pronounced (than on criminal cases): the median disposition time would fall from 13 to 10 months, with the slowest court dropping from 25 to 17 months (excluding the outliers). If we add the optimal reallocation of personnel as well, the median times would further fall to 8 months, and the slowest to about 14 months. For criminal cases, on the other hand, apparently there is a limited effect on the median and slowest court disposition times, which remain respectively at about 5 and 9 months, solely with the break-up of courts or with a 10% increase in personnel and its optimal reallocation.

Last, we consider scenario (iv), i.e., the implementation of best practices and the break-up of bigger courts, together with a 10% personnel increase. The impact of this set of policies is similar to the effect of introducing best practices if we look at median times, but more pronounced when considering the effects on the slowest courts, further reducing disposition times to about 8 months for civil and to 7 months for criminal cases, when combined with the optimal allocation of personnel (bottom panels of Fig. 6).

We conclude by noting that another way of looking at the results presented so far is to calculate the "opportunity costs" of different policy options, in a back of the envelope sort of calculation. We have seen that the most effective policy to reduce disposition times would be the full implementation of best practices: to obtain comparable results personnel would need to be increased by 10% combined with relocation (limited to 5% and 50 km), with the associated costs.[25] Introducing best practices may be cheaper, at least in terms of explicit financial costs, but its implementation would probably be more difficult.

[25] Considering a total number of judges of about 5650, this would correspond to about 565 new judges only. With an initial cost of 70,000 euros per new judge (Senato 2017), this would be an approximate cost of 40 million euros for the first year, only for the judges, to be added to the costs of hiring the complementary administrative staff.

References

Abravanel R, Proverbio S, Bartolomeo F (2015) Misurare la performance dei tribunali. Presentation of March 26, 2015, Osservatorio per il Monitoraggio degli Effetti sull'Economia delle Riforme della Giustizia

Aldashev G (2009) Legal institutions, political economy, and development. Oxf Rev Econ Policy 25(2):257–270

Banker R, Charnes A, Cooper W (1984) Some models for estimating technical and scale inefficiencies in Data Envelopment Analysis. Manag Sci 30:1078–1092

Bartolomeo F (2013) Linee guida sulla revisione della geografia giudiziaria per favorire le condizioni di accesso a un sistema giudiziario di qualità. Discussion paper CEPEJ-GT-QUAL(2013)2, CEPEJ

Bianco M, Palumbo G (2007) Italian Civil Justice's Inefficiencies: a supply side explanation. Mimeo, Banca d'Italia

Bray R, Coviello D, Ichino A, Persico N (2016) Multitasking, multiarmed bandits, and the Italian judiciary. Manuf Serv Oper Manag 18(4):545–558

Buonanno P, Galizzi MM (2014) Advocatus, et non Latro? Testing the excess of litigation in the Italian Courts of Justice. Rev Law Econ 10(3):285–322

Buscaglia E, Dakolias M (1999) Comparative international study of court performance indicators. Washington, D.C., The World Bank

Caponi R (2016) The performance of the Italian civil justice system: an empirical assessment. Ital Law J 2(1):15–31

Carmignani A, Giacomelli S (2009) La giustizia civile in italia: i divari territoriali. Questioni di economia e finanza (occasional papers) no. 40, Banca d'Italia

Carmignani A, Giacomelli S (2010) Too many lawyers? Litigation in Italian civil courts. Temi di discussione (working papers) no. 745, Banca d'Italia

CEPEJ (2016) European judicial systems efficiency and quality of justices (edition 2016). Cepej studies no. 23, European Commission for the Efficiency of Justice, Strasbourg

Charnes A, Cooper W, Rhodes E (1978) Measuring the efficiency of decision-making units. Eur J Oper Res 2(6):429–444

Chen X, Kerstens K, Zhu Q (2021) Exploring horizontal mergers in Swedish District Courts using convex and nonconvex technologies: Usefulness of a conservative approach. Working paper no. 2021-eqm-05

Coviello D, Ichino A, Persico N (2014) Time allocation and task juggling. Am Econ Rev 104(2):609–623

Coviello D, Ichino A, Persico N (2015) The inefficiency of worker time use. J Eur Economic Assoc 13(5):906–947

CSM (2010) Risoluzione concernente la revisione delle circoscrizioni giudiziarie. Rome, Consiglio Superiore della Magistratura

CTFP (2008) La revisione della spesa pubblica. Rapporto 2008. Rome, Commissione Tecnica per la Finanza Pubblica

Dimitrova-Grajzla V, Grajzl P, Sustersic J, Zajc K (2012) Court output, judicial staffing, and the demand for court services: Evidence from Slovenian courts of first instance. Int Rev Law Econ 32:19–29

Draghi M (2011) Considerazioni finali. Technical report, Banca d'Italia

Esposito G, Lanau S, Pompe S (2014) Judicial system reform in Italy—a key to growth. IMF Working paper no. WP/14/32, International Monetary Fund

Felli EL, London-Bedoya DA, Solferino N, Tria G (2008) The "Demand for Justice" in Italy: Civil litigation and the judicial system. In P. F. and R. Ricciuti, editors, Italian Institutional Reforms: A public choice perspective, pp. 155–177. Springer, New York, NY

Kerstens K, Xiaoqing C (2022) Evaluating horizontal mergers in Swedish district courts using plant capacity concepts: with a focus on nonconvexity. Working paper no. 2022-eqm-02

Kittelsen S, Forsund F (1992) Efficiency analysis of Norwegian district courts. J Product Anal 3:277–306

Lewin A, Morey R, Cook T (1982) Evaluating the administrative efficiency of courts. OMEGA Int J Manag Sci 4:401–411

Marchesi D (2003) *Litiganti, avvocati e magistrati. Diritto ed economia del processo civile*. Il Mulino, Bologna

Marchesi D (2008) L'enforcement delle regole. Problemi di efficienza della giustizia civile, riforme intraprese e riforme possibili. I temi dei rapporti trimestrali - Roma, ISAE

Marselli R, Vannini M (2004) L'efficienza tecnica dei distretti di corte d'appello italiani: aspetti metodologici, benchmarking e arretrato smaltibile. Working paper, CRENos

Padrini F, Guerrera D, Malvolti D (2009) La congestione della giustizia civile in Italia: cause ed implicazioni per il sistema economico. Note Tematiche no. 08, Ministero dell'Economia e delle Finanze

Palumbo G, Giupponi G, Nunziata L, Sanguinetti JM (2013) The economics of civil justice: New cross-country data and empirics. OECD Economics Department working papers no. 1060, OECD

Pedraja-Chaparro F, Salinas-Jimenez J (1996) An assessment of the efficiency of Spanish courts using DEA. Appl Econ 28(11):1391–1403

Pereira M, Badin L, Kerstens K, Silva M (2023) Judicial efficiency in Europe: an integrative literature review. Working paper, Catolica Porto Business School

Peyrache A, Zago A (2016) Large courts, small justice! The inefficiency and the optimal structure of the Italian justice sector. Omega 64:42–56

Podinovski VV (2021) Variable and constant returns-to-scale production technologies with component processes. *Oper Res*

Ricolfi L (2009) L'output della giustizia civile: una proposta per misurarlo. Polena 2:73–83

Santos S, Amado C (2014) On the need for reform of the Portuguese judicial system—does data envelopment analysis support it? Omega 47:1–16

Senato Bilancio di previsione dello Stato per l'anno finanziario 2018 e bilancio pluriennale per il triennio 2018–2020. A.s. 2960, Senato della Repubblica (2017)

Silva M (2018) Output-specific inputs in DEA: an application to courts of justice in Portugal. Omega 79:43–53

Svensson L, Färe R (1980) Congestion of production factors. Econometrica 48(7):1745–1753

Tulkens H (1993) On FDH efficiency analysis: some methodological issues and applications to retail banking, courts, and urban transit. J Product Anal 4:193–210

Wei Q, Yan H (2004) Congestion and returns to scale in data envelopment analysis. Eur J Oper Res 153(3):641–660