



Is Sky The Limit? Revisiting 'Exogenous Productivity Of Judges' Argument

By Kamil Jonski* and Daniel Mankowski**

Abstract:

This paper revisits the 'exogenous productivity of judges' hypothesis, laid down in numerous law & economics studies based on the production-function approach. It states that judges, when confronted with growing caseload pressure, adjust their productivity upward, thereby increasing number of resolved cases. We attribute this result to assumptions regarding court productivity. In this paper, we present an alternative – the 'hockey-stick' production function model -- taking into account the time constraints faced by judges. We seek to reconcile this 'production function' model with more traditional and popular approach among practitioners to court performance modelling – the use of weighted caseload methodology. We also propose extended methodology of model evaluation, taking into account its ability to reproduce empirical regularities observed in 'real world' court systems.

Keywords: court efficiency, caseload management, judicial productivity, production function, weighted caseload

JEL Code: H 11, H 50, K40,

Remember that all models are wrong;
the practical question is:
how wrong do they have to be to not be useful?

George E. P. Box

1. Introduction

Regardless of the legal tradition and the level of economic development, any judiciary is confronted with a fundamental dilemma: to effectively administer justice, often under budgetary constraints. One aspect of this dilemma is the challenge of efficient allocation of judges across the court system to ensure the achievement of maximum performance.

The history of contemporary court reform, rooted in a focus on '*law in action*' instead of '*law in the books*', began with works of American legal realists. In his path-breaking address to the annual convention of the American Bar Association in 1906, Roscoe Pound blueprinted the agenda of such reform. As he ascertained, '*Judicial power may be wasted in three ways: (1) by rigid districts or courts or jurisdictions, so that business may be congested in one court while judges in another are idle; (2) by consuming the time of courts with points of pure practice, when they ought to be investigating substantial controversies; and (3) by nullifying the results of judicial action by unnecessary retrials.*' Pound's first point references dilemmas relating to appropriate design of the court system, in order to allocate resources (including judicial work) efficiently.

* Expert, Department of Strategy and Deregulation, Ministry of Justice, Republic of Poland,

e-mail: Kamil.Jonski@ms.gov.pl

** Expert, Department of Strategy and Deregulation, Ministry of Justice, Republic of Poland; University of Warsaw, Faculty of Law and Administration, Institute of Criminal Law, Department of Criminalistics,

e-mail: Daniel.Mankowski@ms.gov.pl

The authors would like to thank J. Beldowski, M. Borowska – Massalska, F. van Dijk, M. Fabri, P. Langbroek, A. Lienhard, B. Ostrom, the participants of 24 seminar 'Reform for Poland' at the Warsaw School of Economics, the participants at 99 'TIGER seminar' of the Kozminski University and the participants of Permanent Study Group XVIII "Justice and Court Administration" during 2013 EGPA Annual Conference in Edinburgh, as well as two anonymous referees and M. Zimmer for their valuable comments and suggestions to the numerous versions of this text.

The views and opinions expressed in this paper are those of the authors and do not necessarily reflect the official policy or position of the Ministry of Justice. Assumptions made within the analysis are not reflective of the position of any Polish government entity.

Half of a century later, Pound's remark that '*dissatisfaction with the administration of justice is as old as law*' was still current. In search for modern solutions for chronic problems, some researchers decided to look at court performance through the lens of quantitative analysis. The seminal study conducted in the late fifties by H. Zeisel, H. Kalven Jr. and B. Buchholz, under the University of Chicago Jury Project – in order to find remedies for the excessive delay in Supreme Court of New York County - significantly influenced modern thinking on court administration. The key contribution made by Zeisel et al. (1959) was the design of statistical methodology focusing on how to estimate judicial staffing requirements essential to reduce case backlogs. It was based on the concept of the so-called '*judge year*' – '*the average workload of one judge during one court year*' (p. 8). A '*judge year*' encompassed judicial capacity, specifying the typical number of cases that can be adjudicated by single judge within one year. If the number of filed and pending cases can be meaningfully translated into '*judge years*', then researchers can begin to devise possible remedies for court productivity delays. Building upon this, Zeisel's team presented their assessment of the magnitude of '*the additional effort required*' to clear the backlog. As they proclaimed - '*The answer is 11.7 judge years*'¹ (p. 59).

Following the initial enthusiasm, the Zeisel team's quantitative approach to judicial staffing was subject to criticism for neglecting the '*demand side*' of the court delay problem. Priest (1989) depicted Zeisel's team approach in terms of '*the metaphor, drawn from the lumber industry, of a logjam. The determinants of the size of the logjam at any point are the rate that logs flow into the lake, the rate that logs flow out of the lake, and the number of logs stuck in the lake from earlier imbalances in the flow.*' Thus, the delay problem or log jam might be addressed by allocating additional judges based on the '*judge-year*' calculus. Alternatively, the '*economic approach*'² – underlying incentives faced by the litigants advocated by Priest (1989) -- '*suggests that (...) plucking logs off the lake increases the rate that logs flow into the lake.*' In other words, although increasing the capacity of the court results in delay reduction, the improvements in court productivity prompts citizens to file more cases because improved court performance reduces litigants' costs. Therefore, court caseloads increase and threaten the productivity gains. Priest dubbed this mechanism '*congestion equilibrium.*'

Despite these criticisms, Zeisel's thinking about court performance fueled the development of the weighted-caseload approach. According to the classic handbook dedicated to the courts as well as legislative bodies (Flango, Ostrom, 1996), weighted caseload methodology³ is applied to translate court caseload (raw number of cases) into workload (amount of time it will take to dispose that caseload) – which then are converted into judicial staffing needs. Since different types of cases require different amount of processing time, analyzing a 'mix' of case filings using weights assigned to particular types of cases, rather than the total (raw) number of filings, offers a better approximation of the workload. In this vein, a '*weighted caseload study is essentially the response to these two questions:*

1. How much judge time, on average, is required to hear each type of case?
2. How much time does a typical judge have available for hearing cases?

In a nutshell, the number of judges required is determined by dividing the amount of judge time needed to hear all cases by the time judges have available to hear cases' (p. 25). Thus, it is just more sophisticated way of thinking about judicial capacity – Zeisel's team '*judge-years.*' Weighted caseload methodology gained popularity in the United States in the 1960s, notwithstanding initial criticism which pointed to its inability to take into account cross-court differences in efficiency (Doyle, 1978). Since then, weighted caseload statistical models have been systematically developed,⁴ for example, to account for productivity differences between metropolitan courts in the largest cities and smaller courts (Flango, Ostrom, 1996, p. 21). They also gained traction outside the USA in continental legal systems like the Netherlands' '*Lamicie-model*' and Germany's '*PEBB\$Y-system*' (Lienhard, Kettiger, 2011).

The logic of weighted caseload systems, built on constant weights⁵ assigned to particular types of cases, drew the attention of law and economics ("L&E") scholars.⁶ Applying a rigorous set of assumptions from mainstream '*rational agent*' economics to their analysis of court case processing, they proclaimed that courtrooms are populated by '*homo oeconomicus.*' Fueled by an array of theoretical and econometric models, researchers endorsed the so-called '*exogenous productivity of judges*' – the hypothesis stating that judges pressured by growing caseloads adjust their case processing activity to increase their productivity – in terms of weighted caseload model – to *squeeze* time required to adjudicate given

¹ Authors themselves called this number '*the cardinal fact of the study and the pivotal figure for the subsequent discussion of remedies* [for backlog] (p.8)

² Also dubbed as '*demand side*' as oppose to the '*supply side*' – focused on court system resources

³ Recommended in the handbook as the best method for assessing judicial needs

⁴ See Tennessee Trial Courts Judicial Weighted Caseload Study (Tallarico et al., 2013) for contemporary example

⁵ In practice, weights are updated every few years – for example in Tennessee first study has been conducted in 1999, second in 2007 and third in 2013 (Tallarico et al., 2013)

⁶ Their credo might be summarized in the statement '*people respond to the incentives*'

case types. Earlier versions of this approach were set forth in less statistically sophisticated US studies: e.g. Levin, 1977; Goerd, 1989.

The contribution of this paper to the literature on quantitative modeling of court performance is threefold. First and foremost, it revisits the '*exogenous productivity of judges*' hypothesis, arguing that its empirical confirmation has been driven by assumptions regarding court's production function -- assumptions offered by L&E scholars. Second, it proposes a '*hockey-stick*' production function that reconciles the '*production function*' approach with weighted caseload methodology. Thus, it enables scholars and practitioners to empirically address typical research questions relating to court case processing efficiency studies with the improved reliability of weighted caseload models. Third, it proposes an extended methodology of model evaluation, rooted in the tradition of '*stylized facts*' popular among macroeconomists, by taking into account the model's ability to reproduce key regularities observable in 'real world' court performance.

The rest of this paper is organized as follows: Section Two briefly reviews law and economics literature devoted to the modeling of court performance and presents in detail the '*exogenous productivity of judges*' hypothesis. Section Three discuss our empirical strategy, and proposes a methodology of model evaluation. Section Four outlines institutional background of the key elements considered in our production function. Section Five presents the simplest formulation of our '*hockey-stick*' production function model. Section Six evaluates the performance of that model, comparing it with Cobb-Douglas models popular in L&E literature. Section Seven offers some insights for further research, incorporating the 'quality dimension' into production-function analysis. Section Eight concludes the discussion.

2. 'Exogenous Productivity of Judges' – The Hypothesis and Empirical Evidence

Although weighted caseload methods dominate the allocation of judicial and staff resources in many jurisdictions, they still attract criticism from academia. As pointed out earlier, weighted caseload methods assume that given types of case require on average measurable adjudication time frames which are converted into case weights. Similar to other occupations, there are limits to how many cases judges, on average, can reliably process in a given time frame. On the contrary, L&E scholars argue that the time required to adjudicate any given type of case can vary, depending on the relative levels of court congestion and the pressure exerted on judges – whether internally or externally – to relieve such congestion. In other words, judges consciously exercise some flexibility in determining how much time is required to adjudicate the case. When overloaded, they may be motivated to simply work faster; their productivity is exogenous. This implies that staffing needs calculated using weighted caseload may result in excess capacity because this judicial flexibility factor is not included in the calculus. Because Beenstock and Haitovsky (2004) offer the most advanced example of such an approach, we rely heavily on it to present and detail the logic and methodological apparatus behind the 'exogenous productivity of judges' hypothesis.

Beenstock and Haitovsky based their argument on two different analytical tools: a theoretical model of the rational judge, and an empirical one – the estimated production function. We present them below in detail.

To explain interdependencies between judicial appointments and the court output, Beenstock and Haitovsky deployed mathematical model of judicial behavior, building upon the influential work of Posner (1993) and Cooter (1983). They assumed that judges (i) are rational '*homo-oeconomicus*,⁷' maximizing their expected utility and (ii) draw utility from that leisure, (Posner, 1993; Cooter, 1983) thus minimizing the effort required for case adjudication. However, they also (iii) draw disutility or negative utility from increasing backlogs, because they may diminish their prestige and prospects for promotion or incentives. Thus the central mechanism behind the model is the trade-off between '*exerting more effort and thereby improving performance, or taking it easier, thereby risking the wrath of the court president*' (Beenstock, Haitovsky, 2004).

A complex analysis of the model, one requiring 'paper and pencil' exercises with differential equations, led the authors to the conclusion that '*Judges are apparently no different to other human beings insofar as they respond to the incentives under which they operate. The amount of judge-time required to complete a case⁸ varies inversely with caseload pressure.*' (p. 368) It means not only that overloaded judges are likely to work in a more efficient fashion but, in addition, that with minimal case processing pressure, judges will protract or expand the time required to finish the case. Thus '*a ceteris paribus increase in the number of judges will tend to reduce the productivity of incumbent judges, because their caseload pressure is reduced. (...). The newly appointed judges naturally increase the output of the court. However, this increase may be partially or even totally offset by the fall in output of the incumbent judges.*' (p. 352) To sum up, according

⁷ The authors of this paper prefer the notion of "*homo-maximizer*" since the objective of the agent (modeled judge) is to maximize discounted expected utility. Certainly, it is not the only approach represented in contemporary (as well as historical) economics, so we believe that notion '*homo oeconomicus*' encompasses also other concepts of 'rationality' e.g. bounded rationality and adaptive learning.

⁸ Case weight in the weighted caseload vocabulary

to the Beenstock and Haitovsky's model there is no clear link between the number of serving judges and a court's overall capacity to handle its caseload.

This theoretical insight has been put under empirical scrutiny. At the initial, descriptive stage, the authors simply plot unweighted caseload numbers against the number of resolutions in each court. Hardly surprising, it turned out that courts with larger caseloads also yielded higher quantities of resolved cases⁹.

Seeking more formal confirmation of their hypothesis, Beenstock and Haitovsky estimated so called 'production function' equation. In economics, the concept of production function is defined as *'the relationship [functional] between the quantity of inputs used to make a good and the quantity of output of that good'* (Mankiw, 2014, p. 376) under available production technology. In the context of court efficiency, Beenstock and Haitovsky linked case processing productivity 'inputs' – namely caseload and number of judges - and 'output' - the number of resolutions - in the following fashion:

$$\ln_resolutions = \alpha + \beta(\ln_caseload) + \gamma(\ln_judges) + \varepsilon \quad (1)$$

As all variables entered the model in logarithms, the production function took the well-known Cobb-Douglas type form¹⁰. Estimation results sealed the 'exogenous productivity of judges' argument, upon which bold evaluative judgment and a landmark policy recommendation have been made: *'Planners of the judiciary might cynically conclude that they should let growing caseload pressure over-work judges even further, thereby increasing their productivity. After all, this strategy has worked over the last 40 years, so why not continue? Better still, if the number of judges makes no difference to the output of the court, why not cut the number of judges?'* (Beenstock, Haitovsky, 2004 p. 368)

Certainly both the existence of a positive correlation between the caseload and the number of resolutions **and** the interpretation that congestion incentivizes judges to work harder have been present in the academic literature long before Beenstock and Haitovsky. Robert Murrell (2001) reviewed such claims and attributed their conclusions to the application of inappropriate statistical techniques like OLS¹¹ regression, incapable of dealing with simultaneous relation between congestion and caseload (see Priest's congestion equilibrium). He also demonstrated, using Romanian courts data,¹² that simultaneous estimation of 'supply' and 'demand' equations - instead of just one 'production function' - can overcome this problem.

Surprisingly, Murrell's estimation approach leading to the *'prosaic conclusion that larger caseloads cause longer delays (p. 2)'* does not resonate within the L&E literature; Beenstock and Haitovsky's Cobb-Douglas production-function approach attracted many followers. Rosales-Lopez (2008) (OLS), deploying the Cobb-Douglas analytical model in Spain's courts, reported that *'a 10% increase in workload produces a 3% increase in judicial output'* (p. 241). Dimitrova-Grajzl et. al (2012) went much further, accompanying similar estimates¹³ with the claims that *'in contrast to judicial staffing, we find that caseload has a statistically significant positive effect on the number of resolved cases.(p. 24)*. The conclusion seems straightforward: *'congruent with Beenstock and Haitovsky's (2004) evidence from Israel, our results therefore suggest that in Slovenia, judge productivity is endogenous in the sense that incumbent judges complete fewer cases in the presence of new judicial appointments'* (p. 24) and *'(...) in Slovenia, an increase in the demand for court services (as proxied by an increase in caseload) incentivizes judges to substantially increase their productivity'* (p. 25).

What is more, (Dimitrova-Grajzl et. al, 2012) disturbingly remarked: *'holding everything else constant, court output does not statistically significantly depend on the number of serving judges...In contrast to judicial staffing, we find that caseload has a statistically significant positive effect on the number of resolved cases.'* (p. 24). This explanation offered by production function econometrics, seems at odds with the common sense assumption that cases are adjudicated by judges. Regrettably, in a subsequent empirical study carried out on judge-level data (Dimitrova-Grajzl et. al, 2012b), the same research team abandoned further exploration of the 'exogenous productivity' hypothesis.

What is noteworthy is that some researchers went beyond the simple Cobb-Douglas equation and estimated more elaborate versions of court production functions. For example Van der Torre et al. (2007) applied the court level translog

⁹ Strong, positive linear correlation between caseload and number of resolutions

¹⁰ The equation (1) can be easily rewritten as: $resolutions = const.*caseload^{\beta}*judges^{\gamma}$. As the expression become linear in logs, researchers are able to use wide range of linear estimators in order to uncover values of β and γ parameters.

¹¹ Ordinary least squares (OLS) - a method for estimating the unknown parameters in a linear regression model

¹² Although imperfect due to data limitations. Despite the fact that Murrell studied commercial courts, he applied **total number of judges** in the examined courts, **including those assigned to non-commercial cases**

¹³ OLS and its instrumental variable extension - two-stage least-squares (2SLS)

production function¹⁴ to Netherlands' data, however squared and interaction variables turned out statistically insignificant, reducing the specification to the well-known Cobb-Douglas form.

3. Setting the Stage - General Remarks on Court's Production Function Empirics

To set the stage for our empirical analysis of interdependences between an individual judge's workload and output, we address some methodological issues which, in our opinion, undermine many earlier empirical inquiries. This section explains our empirical strategy aimed at reducing 'noise' in the data, in order to distill 'the signal' - to follow N. Silver (2012) book's metaphor.

3.1 Model's Consistency

First of all, maintaining the production function approach, we consider it essential to provide conceptual consistency and traceability from known weighted-caseload models or other approaches (e.g. the queuing models – depicting flow of cases through the court in terms of queue¹⁵). In the universe of weighted-caseload model, each judge has a defined docket and time resources which define his or her capacity - the number of cases feasible to adjudicate in a given time period. In contrast, production functions applied in the empirical analysis constitutes 'black boxes' loosely linking (in Cobb-Douglas fashion) aggregated number of cases and judges in order to obtain aggregated number of resolutions. As illustrated in the next paragraph, common data shortcuts only amplify this problem.

For example Mitsopoulos and Pelagidis (2007) in their study of Greek courts' timeliness applied the 'number of total employees that work in any given year for the Ministry of Justice' instead of number of judges in their analysis (p. 224). Barrère et al. (2001) estimated sophisticated stochastic frontier production functions¹⁶ for civil and criminal cases in French first-instance courts. However, in each production function, they included the same total number of judges – despite the fact that some of them have been assigned exclusively to civil while others to criminal cases (Goudriaan, 2003). Similarly, Murrell (2001) - advocating simultaneous estimation of 'supply' and 'demand' equations on the example of Romanian commercial courts - utilized the total number of judges, including those assigned to non-commercial cases. What is more, Cobb-Douglas approach is also error-prone – especially during simulation exercises. In order to illustrate this point, we fed Rosales-Lopez (2008) model with following – hypothetical – data: workload = 550 cases,¹⁷ court staff = 11¹⁸. The model predicted... 560 resolutions! Undoubtedly such result is ridiculous – judges cannot resolve **560** cases, when there are only **550** cases in the docket. In contrast, our proposal offers conceptual traceability in terms of understanding what is going on inside the model and the source of the results.

Thus, our proposal requires distilling, from aggregate data, the amount of judicial time devoted to adjudicate particular type of cases, as well as capturing the number of cases. Contrary to the traditional court level approach,¹⁹ the production function should then link individual workload (and, in more sophisticated version, some individual characteristics) with the number of resolved cases.

In order to practically illustrate this proposal, we limited our empirical inquiry to the first- instance criminal divisions in Polish Regional Courts. The organization of the criminal court system in Poland is defined by a two-instances principle, the right to appeal, and a three-tier hierarchical structure (district courts, regional courts and appellate courts). Generally, the criminal division of district court acts as court of first instance, and the regional court as a court of appeal. However, in the felony cases, the regional court act as court of first instance and the appellate court as a court of appeal (see fig. 1). To improve work organization, the first- and second-instance criminal divisions or regional courts have been separated, and judges are (more or less) permanently assigned to them. Thus, in our dataset, we are able to accurately pick the amount of judicial time (full-time equivalent, hereafter FTE) devoted to adjudicate the pool of felony cases. Since judges assigned to the first instance criminal divisions of regional courts deal exclusively with felonies, we are able to obtain a fairly reliable picture judicial work and case flow.

¹⁴ The translog production function is a generalization of the Cobb–Douglas production function, including interaction variables (and, in this case, squared variables). It is expected to offer more detailed and flexible alternative to the latter.

¹⁵ See RAND (1975) for early review

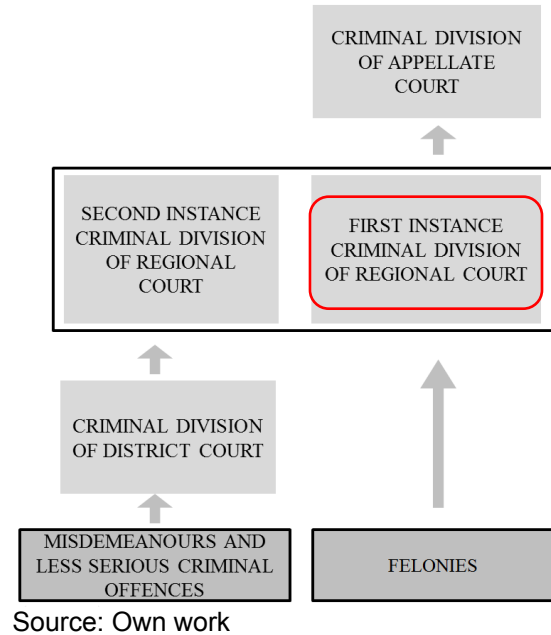
¹⁶ Frontier production function encompass the best performing (efficient) units in the sample – creating so called 'production possibility frontier'

¹⁷ Minimal value of workload in her sample was 1241, thus value 550 seems imaginable in some counterfactual exercise e.g. reform blueprinting.

¹⁸ Values in her sample range from 8 to 11

¹⁹ Typically, one data point represent one court's 'inputs' and 'output' - Dimitrova-Grajzl et. al, 2012b is a notable exception (data points represent individual judges), however they abandoned any attempt to uncover relations between judicial congestion and productivity

Fig. 1. Organization of Polish criminal court system



3.2. Data Aggregation

A common pitfall of much empirical (econometric) research on court performance is a doubtful aggregation of cases. Obviously, cases differ in terms of their complexity. The main types of complexity discussed in the legal literature are (i) legal complexity, (ii) factual complexity, and (iii) participant complexity (Ford, 2013). However, as far as productivity, understood in terms of the number of cases that can be adjudicated in given time period, the notion of 'time-consuming case' seems to be more appropriate than 'complex case'²⁰. While a legally complex case might require a library work and reflection or a flash of genius, a simple case might require extensive amount of paperwork. This consideration is particularly valid in case of heavily procedural frameworks,²¹ requiring judges to perform excessive paperwork.

Typically, researchers either simply aggregate different types of cases or focus on a particular category such as civil, criminal or administrative. To the extent of our knowledge, the studies of Gillespie (1976) and Van der Torre et al. (2007) are notable exceptions because they employed case weights. Unfortunately, up to this date, the Polish courts have not implemented a weighted caseload system. Thus we focused on particular case types, excluding court divisions with a case mix encompassing time-consuming cases – trial-level cases -- and large number of far less time-consuming ones, like writ of payment cases, adjudicated using streamlined procedure by civil and commercial divisions. As we concluded, in civil or commercial divisions, courts with equal *aggregated* caseload might substantially differ in terms of actual workload. In contrast, criminal divisions of regional courts dealing with felonies were a favorable choice because the caseload should provide better approximation of workload. Certainly, focusing on criminal courts cannot solve the problem of data aggregation; however, we expect that it mitigate its impact enough to draw meaningful conclusions.

3.3. Stylized Facts Approach to the Model Evaluation

We also departure from typical approach in terms of model evaluation. In our opinion, standard statistical tests²² are just the tip of the iceberg, and the decisive criterion should be based on the model's ability to replicate empirical patterns observed in the real world. To facilitate such assessment, we follow the so-called 'stylized facts' approach introduced by Kaldor (1961). As Kaldor explained 'Since facts, as recorded by statisticians, are always subject to numerous snags and qualifications, and for that reason are incapable of being accurately summarized, the theorist in my [Kaldor's] view, should be free to start with a 'stylized' view of facts - i.e. concentrate on broad tendencies, ignoring individual detail, and proceed on the 'as if' method, i.e. construct a hypothesis that could account for these stylized facts.' In other words, stylized fact

²⁰ Although it is plausible that complex cases require (on average) more time to adjudicate, the relation between 'complex' and 'time-consuming' cases seems to be far from straightforward.

²¹ See (Djankov et al. 2003) for influential comparative view of procedural formalism as a primary driver of court delay. According to them, formalistic civil law systems rooted in French legal tradition are associated with longer delays as compared with common law systems. See also Siems (2005) for critique of their approach.

²² Carried out during regression model's diagnostics

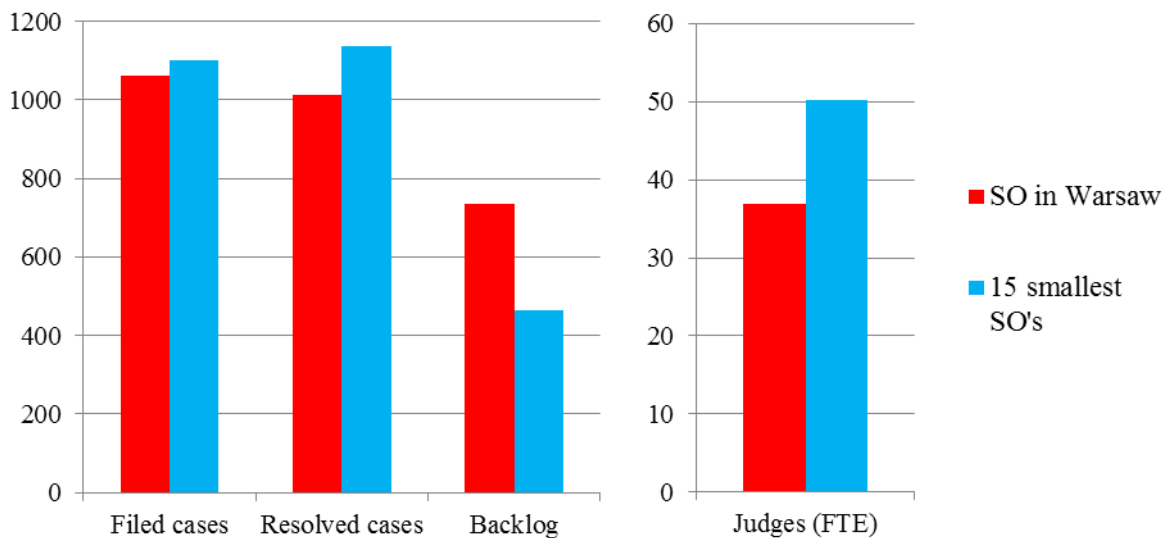
might be described as a widely accepted stable pattern, observed regularity – confirmed by different sources of information: like statistical studies, experience, expert opinion – that demand scientific explanation.²³

We argue that stylized facts can also be formulated in the field of court performance. Thus, any model pretending to reliably depict court performance, including production function models, should be able to replicate them. Below, we list three basic and presumably uncontroversial regularities observed in various court systems:

- There are courts perpetually struggling with persistent and serious backlogs;
- There are courts operating without excessive backlogs which also seem to be able to handle increasing workload without additional resources;
- These courts systematically differ in terms of size and congestion - larger courts in highly populated urban areas and confronted with huge caseloads tend to be sluggish from a case processing perspective and to accumulate increasing backlog.

In our opinion, their universality, confirmed in the literature as well as practical experience, qualifies them as a useful departure point in model evaluation. Figures 2 and 3 briefly illustrate these stylized facts in the context of Polish first-instance criminal divisions of regional courts, henceforth criminal courts. The former compares caseloads, number of resolutions, backlogs and number of judges serving in the largest, in terms of filed cases, court, SO²⁴ in Warsaw, and 15 of the smallest, in terms of filed cases, courts. Despite similar numbers of filings, the Warsaw court resolves fewer cases than the 15 smallest courts, thus it has accumulated a larger backlog. One possible explanation is comparatively smaller numbers of active judges. The latter figure illustrates the third stylized fact; it provides basic statistics depicting the performance of three largest and three smallest, in terms of filed cases, courts. Satellite image of city lights provides insight on population density – illustrating that the biggest and most poorly performing, in terms of backlog, courts are located in the biggest urban centers, while smallest courts are typically located in the less-populated areas of the country.

Fig. 2. Comparison of the filings, resolutions, backlog and the number of serving judges in the largest (Warsaw) and in 15 smallest criminal courts in 2012.

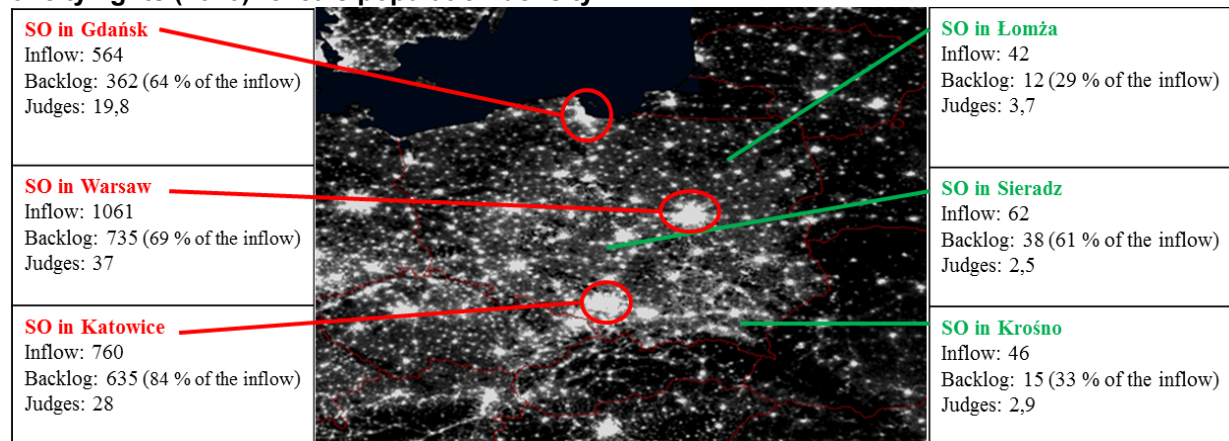


Source: The Ministry of Justice unpublished statistics

²³ On the other hand, N. Kaldor's assessment of broadly agreed patterns has been challenged by other macroeconomists. As remarked by R. Solow (2000) 'It is no doubt that they [Kaldor's facts] are stylized, though it is possible to question whether they are facts (p. 2)'. Despite this critique, stylized facts approach is routinely used in contemporary macroeconomics.

²⁴ Polish abbreviation of Regional Court

Fig. 3. Key data on three largest and three smallest first instance criminal courts. Satellite image of city lights (2010) reveals population density.



Source: www.esa.int, The Ministry of Justice unpublished statistics

4. Institutional Background – Key Elements of the Production Function

In this section we briefly describe key elements considered in our production-function approach, namely criminal courts judges, felony cases and the criminal procedure. Although readers who focus on modelling issues rather than institutional background may want to skip this section, we consider it particularly useful, both for better conceptualization of the model, and to illustrate the complexity of modelled phenomena.

4.1. Criminal Judges

In criminal courts, professional judges sit together with lay judges, serving as non-legal representatives of the local community. Despite strong *de jure* guarantees, lay judges who are chosen by the councils of municipalities within the jurisdiction of a given court tend to be *de facto* primarily inactive ‘observers’ of the adjudication process (Bartnik, 2009).

Although we are unable to provide information about number of lay judges in criminal courts, Table 1 presents the aggregate number of lay and professional judges in the Polish common court system. Substantial decline in the number of lay judges reflects changes in the civil and criminal procedure; the scope of cases requiring lay judge participation has been significantly reduced over past years²⁵. Thus, it seems highly unlikely that the remaining small number of lay judges can be construed as a bottleneck in felony case adjudication. The key determinant of a criminal court’s productivity is the number of professional judges, thus, from our empirical standpoint, lay judges might be omitted without noticeable loose of accuracy.²⁶

Table 1. Professional and lay judges in the whole Polish court system

Years	Number of professional judges (judicial posts stated in the budget)	Number of lay judges
2004 - 2007	9 766 (in 2004) 9 853 (in 2006)	44 000
2008 - 2011	9 890 (in 2008) 10 323 (in 2010 and 2011)	22 000
2012 - 2015	10 348 (in 2012) 10 448 (in 2013 and 2014)	14 000

Source: CEPEJ studies and MoJ unpublished statistics

According to the provisions of Article 28 of the Code of Criminal Procedure (hereafter CCP), for hearing felony cases, *the court shall sit in a panel consisting of one judge and two lay judges*. However, in rare situations, the court may decide to hear the case in a panel comprising three judges because of the *special complexity of the case*. Also, when hearing cases involving offences for which the law provides punishment of life imprisonment, the court shall sit in a panel of two judges and three lay judges. Generally, the share of cases involving atypical panels other than a professional judge and two lay judges should not exceed 5-7% of the caseload. Table 2 illustrates that offences for which the law provides punishment

²⁵ For example they have been excluded from first tier of criminal court system: district courts – see fig. 1

²⁶ Although there is widespread view that lay judges appointment is not a bottle neck in criminal courts, the issue of lay judge change during adjudicating particular case (i.e. death) might be serious obstacle.

of life imprisonment comprise 5-7%. Moreover, as far as special case complexity is concerned, the authors of WC study (see section IV.2) during first half of 2012 report no such cases.

Unfortunately, due to data limitations, we are unable to pick cases requiring atypical panels, thus we have to assume that such situations are uniformly distributed across the courts. Any violation of this assumption will be reflected in the increase of error term (dispersion of the observations around the fitted function).

4.2. The Felony Case

As pointed out in Section Three, felonies adjudicated in criminal courts represent the most serious, and - what particularly important in the context of our modelling exercise – the most time-consuming cases. They usually include an elaborate discovery phase, including witness testimony, as well as drafting complex opinions.

From the legal point of view, these cases are defined in the article 25 § 1 of the CCP as:

- (1) felonies enumerated in the Penal Code (offences for which the law provides for the punishment of 3 or more years) and in special acts (Drug laws and Nazi war crimes),
- (2) misdemeanors specified in Chapters XVI (war crimes) and XVII (treason) and in selected articles (special categories of beatings, euthanasia, abortion, plane hijacking and so on),
- (3) misdemeanors which pursuant to some particular regulation lie within the jurisdiction of the Regional Court (threatening to the journalist and suppressing freedom of the press).

In fact, the first point encompasses vast majority of the cases. The statistics on the *number of adjudicated offenders*, provides some insight to the nature of adjudicated cases.

Table 2. Felonies – share of adjudicated offenders in 2013

Drug offences	15,5 %
Robbery with a firearm, knife etc.	7,3 %
Homicide	4,7 %
Fencing	2,1 %
Rape (selected cases)	1,7 %
Grievous bodily harm	1,5 %
Brawl or a beating with the consequence of death of a human being	1,4 %
Counterfeiting or altering money	1,1 %
Others (not specified)	64,6 %

Source: MoJ unpublished statistics

According to the preliminary and still officially unpublished report from pioneering ‘weighted caseload’ study carried out in 2011 - 2012 (Ministry of Justice, 2013) – herein after (“WC study”) - adjudication of felony cases requires on average ca. 102,1 hours of judicial work²⁷.

Besides felonies, criminal courts handle also so called *cumulative judgments*; these are aggregate penalties imposed on criminal defendants sentenced in multiple judgments rendered by different courts. These cases constitute roughly 33 percent of caseload and are considered as much simpler, requiring 26,3 h of judicial time according to the WC study.

In order to make our results independent from the WC study, since they are rather experimental, not fully operational and officially implemented, we regretfully abandoned the distinction between felony cases and cumulative judgments, assuming that the share of the latter is roughly constant across the courts. We concluded that the ‘average’ criminal case requires 77 hours of judicial time²⁸). Thereby, departure from this assumption translates into the error term in our model. In order to assess plausibility of this assumption, we presented in Appendix I an alternative version of our model, taking *cumulative judgments* explicitly into account. Results obtained in this exercise are in line with baseline model reported in the Section Five.

²⁷ Authors of WC study presented exact figure of 102 hours 04 minutes and 32 seconds. In this paper we rounded up WC Study’s estimates to the first digit.

²⁸Time required to adjudicate ‘average’ criminal case was calculated using felony case weight derived from WC study multiplied by their share in the caseload (102,1 h x 67%) plus cumulative judgment weight multiplied by their share in the caseload (26,3 h x 33%)

4.3. Criminal Procedure

Since the empirical part of this paper employs the production-function approach, using particular functional form to depict relation between inputs (judicial time resources) and outputs (number of resolved cases) under given production technology (the way of transforming inputs into desired outputs), we might think about criminal procedure rules in terms of 'production technology.' This provocative interpretation was also presented in Dusek (2013).

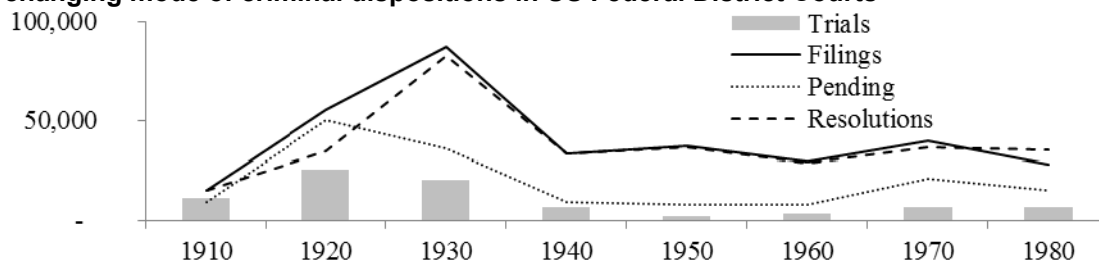
Poland has a continental criminal justice system rooted in the Romano-Germanic law tradition and heavily influenced by the period of communist rule - a consequence of the lengthy Soviet domination after World War II. In 1997, almost a decade after the collapse of communism, a new modern code of criminal procedure (CCP) was promulgated. In line with inquisitorial tradition, it is based on the principle of *objective truth* which requires courts to establish the relevant factual circumstances of the case that constitute the basis for judicial decision (Murzynowski, 2005).

The principle of 'objective truth' affects criminal court's performance in at least two ways. First and foremost, it mandates the judicial obligation to gather the evidence and establish this objective truth as opposed to comparing evidence presented by opposing parties in adversarial proceedings during the discovery, sentencing and opinion-drafting processes, making adjudication more time-consuming.²⁹

Secondly, procedures might affect judicial decision-making process, as analyzed by cognitive psychologists (see Casey et al. 2013 for literature review). For example it seems plausible that inquisitorial, formal procedures of discovery and decision making might constrain heuristic decision making. As noted in (Gigerenzer, Engel, 2006), 'there are a number of restrictions or constraints upon their [heuristics] use. (...) The constraints also differ between the continental and the case law systems' (p. 250). On the other hand, The Royal Commission on Criminal Justice, also known as the *Runciman Commission*, found that in inquisitorial jurisdictions like Germany and France, cognitive bias - judicial tunnel vision - creates greater risk of the wrongful conviction than in an adversarial framework (Leigh, Zedner, 1992). The specific relations between inquisitorial and adversarial frameworks and judicial decision-making exceed the scope of this paper.

Keeping in mind, that according to Beenstock and Haitovsky, 'pressing for compromises, plea bargains and streamlining courtroom hearings' (p. 367) are key mechanisms enabling judges to 'adjust' their productivity, we concluded that the procedural framework of the CCP is an important obstacle to such 'adjustments'. On the contrary, American judges, when confronted with exploding caseload during the prohibition era, improved their productivity by abandoning trials as primary mode of adjudication (see fig. 5). For example, in U.S. state courts, over 90% of felony convictions are routinely obtained by guilty plea.³⁰ In Polish first-instance criminal divisions of regional courts, alternative case-resolution models to the full trial process, which requiring all parties' agreement³¹, account for ca. 20% of all convictions.

Fig. 5. Procedural innovation and court performance – example of prohibition era and changing mode of criminal dispositions in US Federal District Courts



Source: Data from (Heydebrand, Seron, 1990)

5. The Model – 'Hockey-Stick' Production Function

Having in mind the 'exogenous productivity of judges' hypothesis as well as stylized facts approach outlined in Section III.3, we now move toward blueprinting an alternative quantitative model of court performance. Our point of departure is the 'production function' approach; however we also take into account insights from other models, especially the weighted caseload methodology.

²⁹ Anecdotal evidence points to the court's obligation to hear 15 thousand witnesses (repeating roughly the same story, depicted also in the collected financial documentation) in a Madoff-style fraudulent investment scheme case.

³⁰ Felony Defendants in Large Urban Counties, 2009 - Statistical Tables; December 2013, NCJ 243777, Table 21

³¹ Art. 335 and 387 CCP, constituting respectively so called 'conviction without a trial' and 'voluntary submission to penalty during a trial'

Our dataset is based on the yearly statistical reports collected by the Ministry of Justice from particular criminal courts. For each of the 45 criminal courts,³² the dataset contains the amount of judicial time devoted to adjudication (number of judges expressed in FTE's), the number of cases pending from previous year, the number of cases filed in the given year and the number of cases resolved in given year (see table 3 for basic statistics). Since we are unable to examine individual judge's dockets, we calculated – for each criminal court – the 'average judge's' workload and output. The former was calculated as a sum of cases pending from previous year plus cases filed in given year, divided by number of judges (FTE). The latter was calculated by dividing number of cases resolved in the given year by the number of judges (FTE). Thus, our results could be intuitively interpreted in terms of 'average criminal judge's reaction to growing caseload pressure. Figure 6 plots the 'average criminal judge's' workload against output, using data from 44 criminal courts in the 2012 (each data point represents one criminal court).

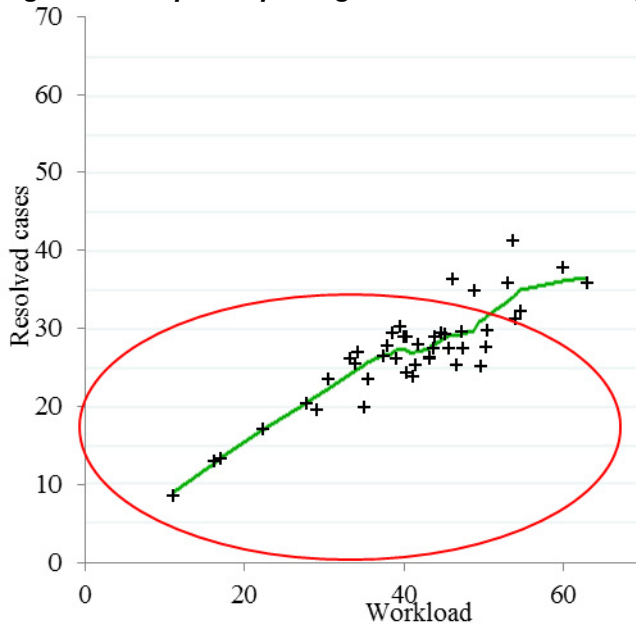
Table 3. Estimation dataset: basic statistics

	Min	Median	Average	Max	St. dev.
Pending cases (from 2011)	12	62	130	688	150
Filed cases	25	144	223	1061	207
Resolved cases	29	147	221	1014	200
Share of cumulative judgments in the caseload	13 %	33 %	33 %	48 %	8 pp
Number of judges (FTE)	1,8	5,7	8	36,9	6,9

Source: Own work

Any parametric estimation of the production function (regression like OLS), inevitably imposes restrictions on the function's shape – in the majority of L&E papers Cobb-Douglas one. Instead, we remain agnostic about all aspects of production-function shape, applying nonparametric regression called LOWESS (*locally weighted scatterplot smoothing*, see Hamilton, 2009, p. 233). Thus, the data is not 'forced' to fit a particular, assumed curve, but directly shapes the estimated line. Green line on the fig. 6 represents LOWESS estimate of the production function.

Fig. 6. Scatterplot exploring behavior of the 'average judge' facing growing workload



Source: Own estimations based on The Ministry of Justice unpublished data

A visual examination of the scatterplot leaves no doubt that there is positive correlation between workload and 'output' of the judge (in line with L&E literature). However, the pattern seems to change with growing caseload. LOWESS results offers a quite blurred but still recognizable picture. Although the relationship appears quite linear, with judge whose workloads are increased able to handle the additional work in line with the 'exogenous productivity' hypothesis, everything changes when the workload exceeds a ca. 38-case threshold. At that point, our 'average criminal judge' has been

³² Court in Olsztyn was omitted as an outlier due to exceptionally high share of cumulative judgments (55,5 % of caseload as compared to the average 33%)

banished from the 'exogenous productivity' paradise; notwithstanding the growing caseload, he or she is able to resolve only ca. 28 cases; his productivity absorption capacity has been exhausted, and the LOWESS production function breaks). Because of this break point between the steep and flatter part of the function, we dub this production function as the 'hockey-stick' model.

Noteworthy, this prosaic story – the existence of a quite rigid limit of cases that might be resolved by a single judge in a finite time period - is in line with the well-established tradition of modeling court performance. Weighted caseload models identify this limit, which corresponds with our break point, as a moment when judicial workload capacity is equivalent available judge-time. Similarly, simulation models of court systems incorporate such mechanisms (see e.g. early works of Taylor et al. (1968) and Stein (1969) or the more contemporary compendium of Nagel (1992, p. 190-200). Also R. Posner, one of the founding fathers of law and economics, assumed that in short run it is impossible to expand court system output (Posner, 2009 p. 128 - 9). Even Beenstock and Haitovsky themselves foresee the existence of 'some upper limit of coping' (p. 366) – however abandoned its further exploration. Recent literature on cognitive psychology of judicial decision-making also offers strong support for the case of 'upper limit of coping' – see for example (Klein, Mitchell, 2010). Going back to the 'stylized facts' developed in the previous section, one can see that 'hockey-stick production function' – despite its simplicity – offers a potential to replicate them. Smaller, less congested courts - operating below break-point, are able to adjust their productivity to growing caseload (until they reach the threshold). On the other hand, larger and often congested courts with more than ca. 38 cases in the judicial docket, are typically unable to further increase their productivity, thus backlog growth is inevitable without the provision of additional judicial resources.

Indeed, the empirical analysis presented in this paper, although crude if compared with reviewed empirical approaches, tells a quite reasonable story backed up by consistent combination of the agnostic, data driven analysis and common sense. In the next section we propose a more rigorous framework to respond to Box's question about usefulness of our model.

6. Assessing Calibrated 'Hockey-Stick' Judicial Production Function

Since court delay, the main motivation behind empirical modeling of court performance, is a dynamic phenomenon, usually deteriorating over time, any model claiming practical application should be able to replicate the process of backlog accumulation. In order to verify our model through this lens, we performed simple counterfactual exercise attempting to visualize whether it is able to replicate actual dynamics of criminal courts 'output' and backlog during 2008 – 2012 period³³.

For the purpose of this exercise, we calibrated – on the basis of fig. 6 - the simplest possible *hockey-stick production function* formula (see green dashed line in fig. 7):

$$av. judge output = \begin{cases} 0.75(av. judge workload), & 38 < av. judge workload \\ 28, & av. judge workload \geq 38 \end{cases} \quad (2)$$

The mechanism of our counterfactual exercise is straightforward – for each court we obtained **actual data** on initial backlog (from 2007) and the number of filed cases, as well as the number of serving judges (FTE) in each consecutive year. Then we applied our model (equation 2) to calculate **artificial** time series for number of resolved cases and corresponding backlogs during a five-year horizon. We then assessed model performance, comparing this artificial time series with real-world data. Secondly, we carried out an analogical exercise with constant returns from the scale Cobb – Douglas production function³⁴ (see red dotted line in fig. 7). Taking into account 'stylized facts' listed in the subsection III.3, we selected the three largest and three smallest criminal courts, in terms of filed cases, as our test site (see fig. 3 for basic data on their performance).

Results are plotted in fig. 8. As can be seen, a very simple model offers a reasonably useful explanation of court productivity as well as backlog dynamics in the smallest and largest courts in mid-term horizon. We consider these results very promising.

In our opinion fig. 8 offers also tough reality check for competing model, estimated under assumption that a court's 'production function' has a constant returns of scale Cobb-Douglas form. Although the results for the smallest courts are comparable, the Cobb–Douglas model visibly fails to replicate the process of backlog accumulation in three largest courts. In line with hypothesis that overloaded judges increase their productivity, the model systematically overstated the number

³³ The period was selected due to comparable data availability

³⁴ In case of variable returns of scale Cobb-Douglas production function, backlog in the biggest courts would be attributed to their size.

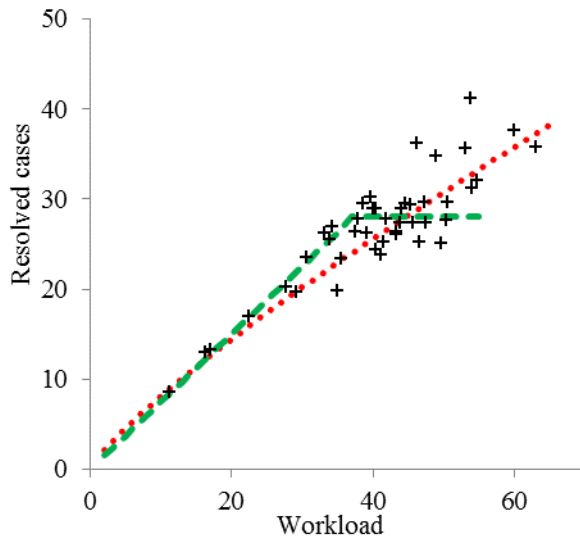
of resolved cases, thereby predicting systematically declining backlog. One might bitterly conclude, that in the universe of such model, the backlog simply clears itself, absent any need for any backlog-reducing policies.

What is particularly important, as noted by Gillespie (1976), who found evidence that in US federal district courts, judicial response to the growing caseload differs between courts operating 'below' and 'over' their capacity;³⁵ 'until resources are fully employed, productivity is not relevant, as the additional costs of meeting increased demand are zero' (p. 259). Also New Zealand's Report of the Ministry of Justice (2004) found evidence in favor of a nonlinear relation between court's 'inputs' and 'outputs'. In line with this reasoning, measuring court efficiency in the Cobb-Douglas framework, thus ignoring the 'upper limit of coping' might lead to the misattribution of inefficiency – e.g. to the court size.

The second test, aimed at assessment and verification of our *hockey-stick production function*, addresses its 'microfoundations'. Since the break point on the function represents the situation when 'average criminal judge' reached his capacity, it conceptually corresponds with a situation when the judicial workload is equivalent to the judge year in the weighted caseload model. Thereby, points unveiled by both methods should coincide. In this case we are able to conduct a falsification exercise – if our production function is not a mere mirage, the break point should hit the judicial capacity derived from WC study.

The so-called 'process approach' applied in the WC study emphasized criminal procedure design (for example case reception has been divided into eight possible judicial activities). The authors carefully analyzed CCP, identifying almost 100 steps required to fully process felony cases. Then, using sample of 430 felony cases and 367 cumulative judgments, they assessed the time required to complete particular steps and their frequency, calculating an aggregated criminal case weight. Such approach provides a genuine link between procedural rules, the time required to adjudicate felony case and judicial capacity. On the contrary, our framework provides only a bird's eye view of these links (as painted by aggregate, court-level data). Figure 9 confronts our LOWESS estimates with the value equalizing judicial workload and judge-year, derived from WC study.

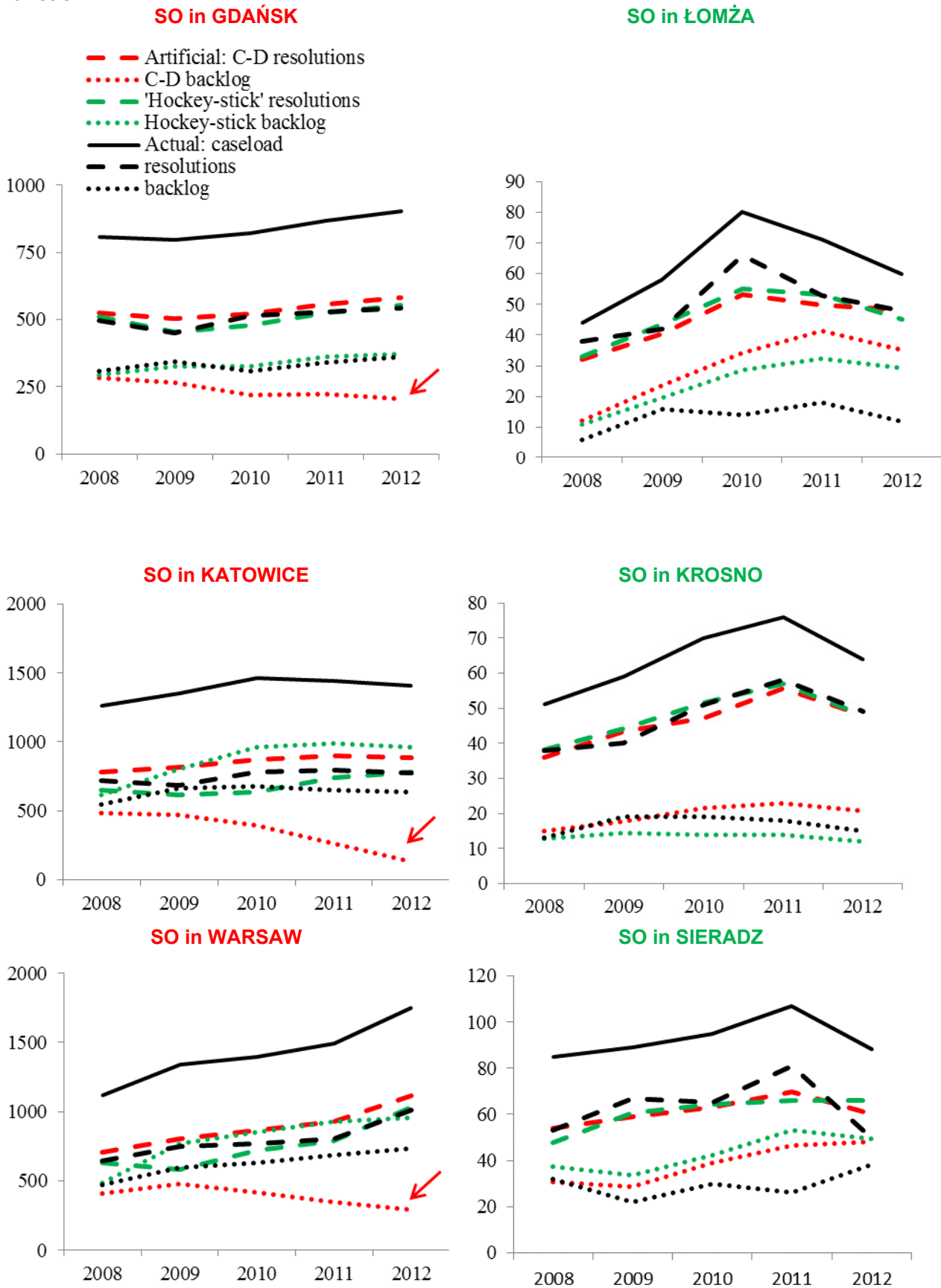
Fig. 7. Production functions employed in counterfactuals



Source: Own estimations based on The Ministry of Justice unpublished data

³⁵ According to Gillespie, even courts operating 'over' their capacity has been able to absorb some additional caseload (however to a lesser extent than in those operating 'below' capacity) – as he concluded 'Courts - even those at 'capacity' - retain the ability to process additional demand with existing resources'. Discussion on procedural framework in the context of court production function, presented in subsection IV.3. offers potential explanation for this finding.

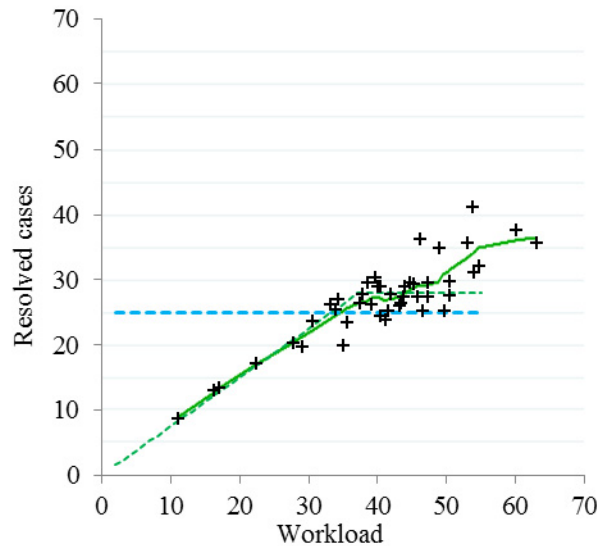
Fig. 8. The results of counterfactual exercise –hockey-stick vs. Cobb - Douglas production function



Source: Own estimations based on The Ministry of Justice unpublished data

This graphic demonstrates that despite a pyramidal gap in the sophistication and costs of both applied methods, they offer fairly consistent estimates of judicial capacity. Undoubtedly, it is quite powerful argument in favor of proposed formulation of judicial ‘production function.’ Since this very basic model, omitting other production factors such as courtrooms,³⁶ obvious controls like case complexity and various sources of technical inefficiency, is able to deliver pretty reasonable results, it offers substantial potential for further development.

Fig. 9. Limits for judicial capacity – hockey-stick break point (green line: LOWESS estimate - solid, equation 2 – dashed) and WC study judicial capacity (blue dashed line)



Source: Own estimations based on The Ministry of Justice unpublished data, and WC study

Given promising results described above, Bayesian estimation of the hockey-stick production function seems to be a natural extension of our work. It would enable rigorous combination of *prior* knowledge (case weights derived from separate studies, stemming from the detailed consideration of procedure framework) and aggregated data - in order to fit judicial production function. Such models would enable e.g. systematic *ex ante* impact assessments of quite detailed procedural reforms.

7. Judicial Productivity and Quality

Traditionally, research papers on court performance, building upon production function approach and usually written by economists, focus entirely on quantitative aspects like relation between filed and resolved cases or time to disposition.³⁷ While this paper is also almost exclusively devoted to the quantitative modelling of judicial productivity, we decided to include some ‘*qualitative aspects*,’ in order to better illustrate the analytical potential of proposed modelling tool.

Undoubtedly, measuring the ‘quality’ dimension of judicial performance is challenging, both in defining the quality measurement standards and in ensuring consistent application of those standards. Traditional measures applied in empirical studies build upon higher court reversals or number of citations (Posner, 2000) – as a ‘*proxies for quality of judicial output*’ (p. 711). Other approaches stress the importance of so called ‘procedural justice’ (Hough et al., 2013) – concepts like trust, neutrality and voice (all of them difficult to precisely define). Also practitioners struggle to develop a set of balanced and realistic performance measures covering aspects like *access and fairness*, inextricably linked with quality of judgments.³⁸

Unfortunately, due to paucity of data, we are unable to design and to implement such quality measures and to integrate them into our empirical framework. However, thanks to Court Watch Poland- an NGO ‘*supporting positive changes in the*

³⁶ Anecdotal evidence suggests that lack of courtrooms is not the primary concern of criminal judges, especially in the largest regional courts – thus such simplification seems to be justified in the presented case.

³⁷ Rosales-Lopez (2008) is a notable exception, as she considered ‘*quality of resolutions*’ using reversal rate.

³⁸ E.g. CourtTools designed by the National Center for State Courts, <http://www.courttools.org/Trial-Court-Performance-Measures.aspx>. Also Trial Court Performance Standards, developed by U.S. Bureau of Justice Assistance, offers quite interesting approach: its *Measure 3.1.1 requires panels of expert practitioners to assess the court’s adherence to legal requirements by examining documents, case files, and other court records.*

*Polish system of justice through citizen court monitoring*³⁹- we are able to incorporate some 'quality' dimensions to our production function framework. Court Watch Poland generated the dataset encompassing several criminal courts, building upon 'citizen trial monitoring' – observation of the court sessions carried out by trained⁴⁰ volunteers. The methodology applied by the Court Watch Poland, based on forms filled in courts, was outlined in (Pilitowski, Burdziej, 2013).

Thus we are able to match our production function (indicating overloaded courts) with data derived from 'citizen trial monitoring' – thereby visualizing potential trade-offs. To facilitate comparability with our production function dataset, we selected 378 criminal trials observations carried out in 12 criminal courts in 2012⁴¹.

Fig. 10 plots calibrated 'hockey-stick' production function (eq. 2) and data points representing courts covered in Court Watch Poland dataset⁴² encompassing:

- Judicial lateness (arising from judicial fault⁴³). If the proportion of court sessions delayed as a consequence of judicial fault exceeds 10 percent of observations, the data point is marked red;
- Judicial politeness. Volunteers recorded whether the judge addressed someone in a rude or aggressive manner - if so data point was marked red. Although highly subjective,⁴⁴ this measure seems very interesting, since irritation is a well-recognized symptom of stress (see Arnetz, Ekman (2006), especially the Chapter 'Work, Stress and Productivity'),
- The mode of preparing minutes. Typically, judges dictate aloud the text of the minutes. When the judge corrected or amended the minutes prepared by the court clerk during more than 5% of the observed sessions, the data point was marked red. This measure is also quite interesting, in terms of judicial work organization and cooperation with the support staff.

Although data plotted in figure 10 are only illustrative and by no means should be interpreted as conclusive empirical evidence (due to small sample and subjective nature of applied measures) – there appears to be some trade-off between increased productivity and quality. For example, judicial lateness and irritation tends to be more likely around and beyond the break point of judicial capacity, which is consistent with insights from psychological studies of stress in work environment (Arnetz, Ekman, 2006).

Moreover, similar effects might influence judicial decision-making. As suggested in (Klein, Mitchell, 2010) it seems reasonable 'to predict more intuitive processing by judges facing greater workload pressures and more deliberation by those who are relatively unconstrained by such pressures (p. 38).' Undoubtedly, the interdependences between judicial workload, procedural framework (adversarial or inquisitorial) and modes of decision-making (fast and frugal heuristics,⁴⁵ cognitive biases, deliberation) offer promising areas for further research.

³⁹ <http://courtwatch.pl/>

⁴⁰ 'During the (...) training participants acquire basic information on the Polish court system, including rules for conduct in the courtroom and citizen rights (depending on the level of knowledge of the participants). This is expected to make them more confident without losing the "ignorance" characteristic for those who have never dealt with courts before' (Pilitowski, Burdziej, 2013 p. 21)

⁴¹ In the smallest court in Przemyśl (25 cases filed in 2012) there were only 6 observed trials, while in the largest Warsaw (1061 cases filed in 2012) 37. Surprisingly, in Kraków (416 cases filed in 2012) there were 106 observed trials. Thus, as stressed at the end of this section, presented results are only illustrative and by no means should be interpreted as conclusive empirical evidences.

⁴² The authors would like to thank B. Pilitowski and S. Burdziej for providing the 'Citizen Trial Monitoring' raw database

⁴³ Situations when judge was late because his previous session extended have been removed, as we concluded that there is high probability that they are independent from the judge. Certainly such cases might also indicate judicial inability to exert control over the courtroom. Robustness check using total share of sessions started late, yield qualitatively similar results. Court Watch Poland's volunteers need to record judicial lateness (yes or no, if yes – its duration in minutes) and its reason (e.g. extended previous session) thus the measure seems to be quite objective.

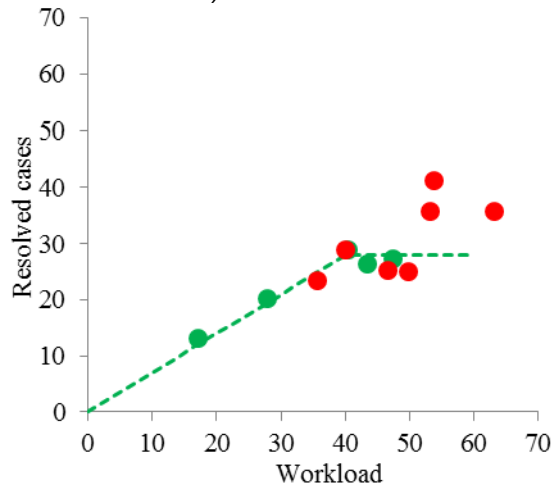
⁴⁴ Note that such assessment is inherently subjective and involves observer's personal sensitivity – thus should be interpreted with caution. Moreover, the 'citizen trial monitoring' has been typically conducted by volunteer law students, typically engaged in other human rights promoting activities. Thus one might reasonably expect them to be biased (overly critical towards judges) and, despite training, unaware of daily courtroom practice. As acknowledged in (Pilitowski, Burdziej, 2013) 'Their [volunteers] assessment of the situation will inevitably be subjective (and not always accurate!), but citizens also experience the court subjectively (p. 14)'

⁴⁵ See (Gigerenzer, Todd, 2000) for discussion on fast and frugal heuristics - simple rules for making decisions when time is pressing and deep thought an unaffordable luxury (so called ecological rationality)

Fig. 10. Judicial production function and work performance observed during the ‘trial monitoring’

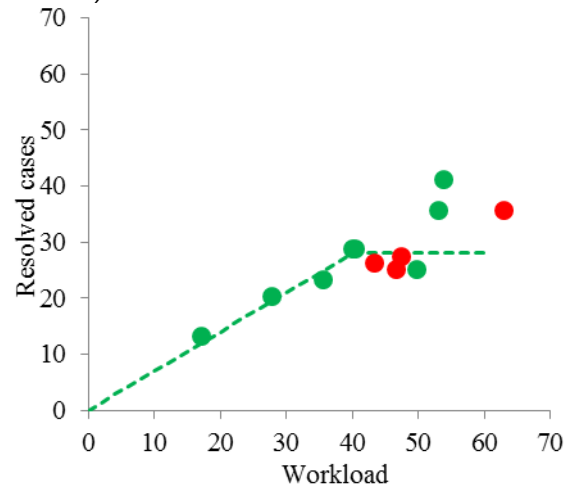
Judicial lateness

(red points denote courts where share of sessions delayed from judicial fault exceeds 10 % of observations)



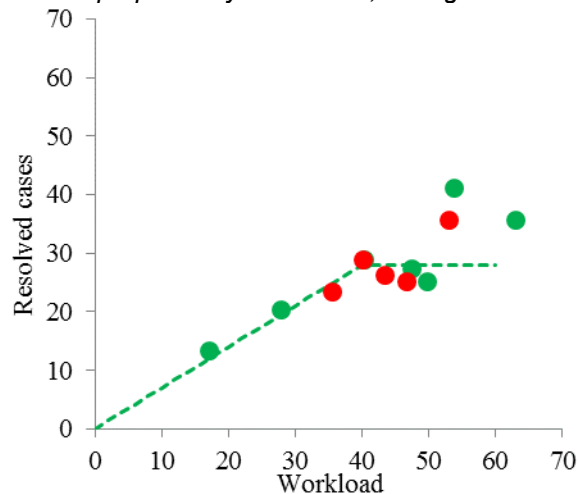
Judicial politeness

(red points denote courts where judge at least one time addressed someone in a rude or aggressive manner)



The mode of preparing minutes

(red points denote courts where judge only corrected or complemented minutes prepared by court clerk, during more than 5 % of observed sessions)



Source: Own calculations based on The Ministry of Justice unpublished data and Court Watch Poland dataset

8. Conclusions

Over one hundred years after Roscoe Pound’s speech, reformers across the world still try to cope with delayed (thus denied) justice, improving methods of assessing judicial staffing needs, optimizing structure of judicial districts and streamlining procedures. The work associated with such reform efforts can be supported with court system models which might be applied to generate counterfactuals, perform stimulations and impact assessments that allow reform designers to verify their prescriptions in the artificial world, instead of imposing experiments on the living tissue of the court system.

Modeling court performance is a specific endeavor, too fluid to be a science and too rigorous to be an art. Although models by definition constitute simplified, ‘alternate universes,’ to remain useful they need to properly capture the basic characteristics of our own universe. Otherwise, obtained answers might be spurious and formulated recommendations misguided.

This paper provides evidences that so called 'exogenous productivity of judges' hypothesis might be attributed to a modelling flaw, originating in neglecting time constraints faced by judges or in *ad hoc* application of specific functional form, without in-depth consideration of their plausibility. Thus our primary contribution to the existing literature on court performance, particularly employing production functions, is the proposal for alternative formulation of court's production function, taking into account time constraints faced by judges. The proposed 'hockey-stick' production function reconciles the weighted-caseload approach, popular among practitioners on both sides of the Atlantic, with production function approach advocated by L&E scholars. Thus it enables researchers to address research questions typical for production function literature (i.e. comparing court performance, identifying drivers of court's inefficiency and proposing optimal allocation of judges) without losing consistency and traceability of weighted caseload methods.

Building upon this toolkit, we demonstrated the common sense conclusion: contrary to the views promulgated by the prophets of the 'exogenous productivity,' judges operate in the universe where adjudication takes time and a day lasts no more than 24 hours. Thus, while appreciating the importance of technical efficiency improvements as well as potential pitfalls of performance measurement, 'planners of the judiciary' must realize that judicial productivity is objectively constrained. When these constraint is met, additional staffing or procedural streamlining is necessary in order to avoid backlog explosion.

Another contribution of this paper is the proposed extended method of model evaluation, rooted in 'stylized facts' approach known from contemporary macroeconomics. It requires the model to be able to replicate key regularities observed in the real world, not just pass the battery of statistical tests.

We also believe that our results offer some value to practitioners – especially as a cheap and simple tool for evaluating weighted caseload studies (like in exercise depicted in the fig. 9).

Last but not least, approach presented in the Section Seven might offers valuable insight for further research on judicial productivity, the impact of work stress and the cognitive psychology of judicial decision-making in order to encompass all aspects of adjudication.

9. References

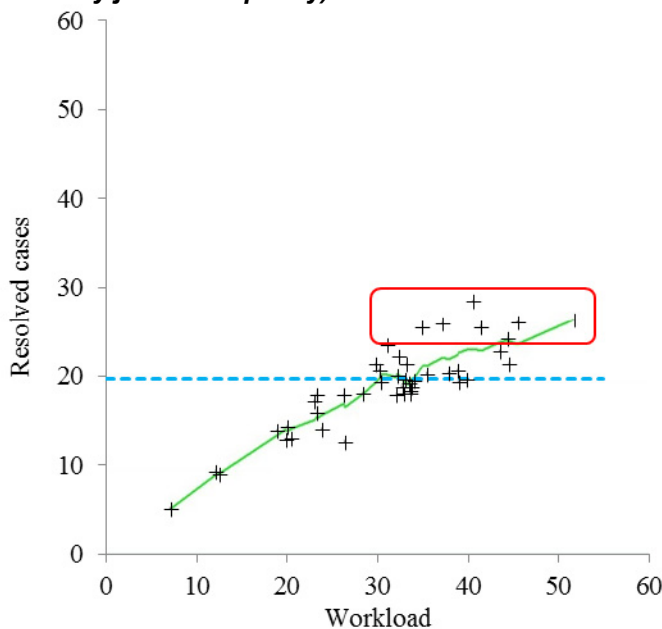
- Arnetz B., Ekman R. (2006)** *Stress in Health and Disease*, Wiley-Blackwell
- Barrère C., Gallien J.-M., Wilmotte C., Kobielski J., Bescont L., Come T., Facchini F., Harnay S., (2001)**, *La Qualité de la Justice: Une Analyse Économique Exploratoire*, Reims: Université de Reims Champagne-Ardenne (Centre d'Études et de Recherches en Analyse de Systèmes).
- Bartnik A. (2009)** *Judge or observer : the role of lay assessors in the justice administration of the Polish Third Republic*, TRIO
- Beenstock M., Haitovsky Y., (2004)**, *Does the appointment of judges increase the output of the judiciary?*, International Review of Law and Economics, Elsevier, vol. 24(3), pages 351-369, September.
- Casey P., Burke K., Leben S., (2013)** *Minding the Court: Enhancing the Decision-Making Process*, International Journal For Court Administration, February 2013
- Cohen A., Harcourt G. (2003)** *Whatever Happened to the Cambridge Capital Theory Controversies?* Journal of Economic Perspectives—Volume 17, Number 1—Winter 2003—Pages 199–214
- Cooter R., (1983)**, *The Objectives of Private and Public Judges*, Public Choice, Vol. 41, No. 1, pp. 107-132
- Dimitrova-Grajzl, 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*, International Review of Law and Economics, Vol. 32, No. 1,
- Dimitrova-Grajzl, V., Grajzl P., Sustersic J., Zajc K., (2012b)**, *Judicial Incentives and Performance at Lower Courts: Evidence from Slovenian Judge-Level Data*, Review of Law and Economics, Vol. 8, No. 1, pp. 215-252,
- Djankov S., La Porta R., Florencio Lopez-de-Silanes F., Shleifer A. (2003)** *Courts*, The Quarterly Journal of Economics (2003) 118 (2): 453-517
- Doyle J., (1978)**, *Comparing Court Productivity*, JUDICATURE Volume: 61 Issue: 9 (April 1978) , pp.: 416-421
- Dusek L. (2013)** *The effects of simpler and faster criminal procedure on criminal case outcomes: Evidence from Czech district courts*, available at: <http://kie.vse.cz/wp-content/uploads/fastprocpaper02-3.pdf>
- Flango V., Ostrom B., (1996)** *Assessing the Need for Judges and Court Support Staff*, National Center for State Courts
- Ford S. (2013)** *Complexity and Efficiency at International Criminal Courts*, Available at SSRN: <http://ssrn.com/abstract=2370540>
- Gigerenzer G., Engel C., (2006)** *Heuristics and the Law*, MIT Press
- Gigerenzer G., Todd P., (2000)** *Simple Heuristics That Make Us Smart*, New York: Oxford University Press
- Gillespie R. (1976)** *The Production of Court Services: An Analysis of Scale Effects and Other Factors*, Journal of Legal Studies: Vol. 5: Iss. 2, Article 4

- Goerdt J., Lomvardias C., Gallas G., Mahoney B. (1989)** *Examining Court Delay. The Pace of Litigation in 26 Urban Trial Courts, 1987*. Williamsburg: National Center for State Courts
- Goudriaan R., Halbersma R.S., Kerstens K., Snijder J.K., Thio V., (2003)**, *Rechtspraak en productiviteit Een internationale verkenning*, Advies voor de Raad voor de Rechtspraak
- Hamilton L., (2009)**, *Statistics with STATA: Version 10*, Cengage Learning
- Hough M., Jackson J., Bradford B. (2013)** *Legitimacy, Trust and Compliance: An Empirical Test of Procedural Justice Theory Using the European Social Survey*, [in] Tankebe J., Liebling A. (eds.) *Legitimacy and Criminal Justice: An International Exploration*, Oxford: Oxford University Press
- Kaldor N. (1961)** *Capital Accumulation and Economic Growth* [in] F.A. Lutz and D.C. Hague, eds., *The Theory of Capital*, St. Martin's Press
- Klein D., Mitchell G., (2010)** *The Psychology of Judicial Decision Making*, Oxford University Press
- Leigh L., Zedner L., (1992)** *A report on the administration of criminal justice in the pre-trial phase in France and Germany*, Royal Commission on Criminal Justice
- Levin M. (1977)** *Urban Politics and the Criminal Courts*, University of Chicago Press, Chicago
- Lienhard A., Kettiger D., (2011)**, *Research on the caseload management of courts: methodological questions*, Utrecht Law Review, Volume 7, Issue 1, January 2011
- Mankiw N., (2014)** *Principles of Economics, 7th Edition*, Cengage Learning
- Ministry of Justice (2013)** *Final Report: Research and recommendations on workload and work standards for all professional groups in the court system (PL: Przeprowadzenie badania i opracowanie raportu wraz z rekomendacjami: analiza obciążenia pracą i opracowanie standardów pracy dla wszystkich grup zawodowych w sądownictwie, Raport Podsumowujący)* Retrieved from: <http://kurator.webd.pl/wp-content/uploads/2014/01/Raport-Podsumowuj%C4%85cy-Zadania-4.pdf> 20 Aug. 2014
- Ministry of Justice, New Zealand (2004)** *Report of the Ministry of Justice - Baseline Review*, Retrieved from: <http://www.justice.govt.nz/publications/global-publications/r/report-of-the-ministry-of-justice-baseline-review>
- Mitsopoulos M., Pelagidis T., (2007)**, *Does Staffing Affect the Time to Dispose Cases in Greek Courts?*, *International Review of Law and Economics*, Vol. 27, No. 2, 2007
- Murrell P., (2001)**, *Demand and Supply in Romanian Commercial Courts: Generating Information for Institutional Reform*, Available at SSRN: <http://ssrn.com/abstract=280428>
- Murzynowski A. (2005)** *Criminal Procedure* [in] Frankowski S., Bodnar A. (2005) *Introduction to Polish law*, Kluwer Law International
- Nagel S., (1992)**, *Computer-Aided Judicial Analysis: Predicting, Prescribing, and Administering*, Praeger
- Ostrom B. et al., (2003)** *Minnesota Judicial Workload Assessment, 2002*
- Pilitowski B., Burdziej S., (2013)** *Court Monitoring Methodology*, Court Watch Poland Foundation, Retrieved from: http://courtwatch.pl/wp-content/uploads/Court_Monitoring_Methodology.pdf
- Posner R., (1993)**, *What Do Judges Maximize? (The Same Thing Everybody Else Does)*, *Supreme Court Economic Review*, Vol. 3., pp. 1-41.
- Posner R., (2000)**, *Is the Ninth Circuit Too Large-A Statistical Study of Judicial Quality*, *The Journal of Legal Studies*, Vol. 29, No. 2 (June 2000), pp. 711-719
- Posner R., (2009)**, *The Federal Courts: challenge and reform*, Harvard University Press
- Pound R., (1906)**, *The Causes of Popular Dissatisfaction with the Administration of Justice*, Address Before the Annual Convention of the American Bar Association (Aug. 29, 1906)
- Priest G. (1989)**, *Private Litigants and the Court Congestion Problem*, Faculty Scholarship Series. Paper 613.
- RAND (1975)** *Criminal Justice Models: an Overview*, RAND Corporation
- Robel L. (1990)** *Caseload and Judging: Judicial Adaptations to Caseload*, *Brigham Young University Law Review*, 3 (1990)
- Rosales-Lopez V., (2008)**, *Economics of Court Performance: An Empirical Analysis*, *European Journal of Law and Economics*. N° 25. Pg: 231-251.
- Siems M. (2005)** *Numerical Comparative Law: Do We Need Statistical Evidence in Law in Order to Reduce Complexity?*, *Cardozo Journal of International and Comparative Law*, Vol. 13, pp. 521-540
- Silver N. (2012)** *The Signal and the Noise: Why So Many Predictions Fail — but Some Don't*, Penguin Books
- Solow R., (2000)** *Growth Theory: An Exposition*, Second edition, Oxford University Press
- Stein S., (1969)**, *Computer Simulation and the Problem of Court Congestion*, 27(1) *University of Toronto Faculty of Law Review*. 94
- Tallarico S., Douglas J., Friesset E. (2013)** *Tennessee Trial Courts Judicial Weighted Caseload Study, Final Report*, National Center for State Courts
- Taylor J., Cohen R., (1968)**, *Data analyses and simulation of the District of Columbia Trial Court system for the processing of felony defendants*, Institute for Defense Analyses United States of America
- Van der Torre A., Jonker J., van Tulder F., Steeman T., Paulides G., (2007)**, *Rechtspraak: productiviteit in perspectief*, Zeisel H., Kalven Jr. H., Buchholz B., (1959-), *Delay in the Court: an analysis of the remedies for delayed justice*

APPENDIX I

In order to ensure that our results are fully independent from WC Study, up to this point we employed its results only to evaluate our model. Since reader might be concerned with two assumptions made in section IV, namely that: (i) criminal court's capacity is driven by the number of regular judges, so lay judges might be excluded from the model, and that (ii) *cumulative judgment* cases are uniformly distributed across the criminal courts. While the former seems uncontroversial in light of practical experience as well as data presented in the table 1, the latter is more problematic. Despite the exclusion of the court in Olsztyn (due to disproportionately high number of *cumulative judgments*), the diversity between remaining 44 courts remains substantial (see table 3). Thus, building upon WC Study weights, we recalculated our model, assigning to each felony case value 1, and for each cumulative judgment 0,26 (26,3h divided by 102,1h). The results (including court in Olsztyn) are plotted on the figure A.1. Again, *LOWESS* estimate of production function breaks in the point corresponding with judicial capacity derived from WC Study. Moreover, ca. 25 felony cases seems to constitute glass ceiling for the judicial productivity – holding for 30 as well as 50 cases in the docket. Thus, it confirms the main finding of this paper, namely that judicial production function is 'hockey-stick' shaped – with break point between steep (from 0 to the full capacity) and more flat (full capacity and beyond) part.

Fig. A.1. Hockey stick production function - (green line - *LOWESS* estimate, blue dashed line - *WC* study judicial capacity)



Source: Own estimations based on The Ministry of Justice unpublished data

