ei estudos internacionais • Belo Horizonte, ISSN 2317-773X, v.6 n.3 (2018), p.82 - 97

Models, explanation, and the pitfalls of empirical testing

Modelos, explicação e as armadilhas dos testes empíricos

DOI: 10.5752/P.2317-773X.2018v6.n3.p82

Enzo Lenine[1]

[1] Enzo Lenine Nunes Batista Oliveira Lima is Professor of International Relations at the University of International Integration of the Afro-Brazilian Lusophony (UNILAB-Malês), Brazil. His research interests are mostly connected to methodology, formal models and rational choice theory, and hierarchies of knowledge. He has recently published a bibliometric analysis in the International Political Science Review as part of his work on the hierarchies of knowledge in the discipline.
E-mail: lenine@unilab.edu.br
Salvador/Brazil.

ORCID: 0000-0001-5280-4252

Submitted in march 25, 2018
Accepted in may 22, 2018

Abstract

Formal models constitute an essential part of contemporary political science and International Relations. Their recent history is tightly tied to the developments of Rational Choice Theory, which is considered to be the only deductive theory in the social sciences. This unique character, especially its manifestation through mathematical symbolisms, has caused profound schisms and criticisms in the discipline. Formal models have constantly been accused of being built on unrealistic assumptions of human behaviour and social structure, rendering as a result either trivial predictions or no empirical prediction at all. Nevertheless, these criticisms frequently ignore essential elements of the concept of explanation and how it is applicable to formal modelling. In this paper, I provide an approach to mathematical modelling that considers the challenges of designing and performing empirical tests of predictions generated by formal models. Rather than disqualifying or falsifying models, empirical tests are paramount to the tailoring of more grounded explanations of political phenomena and should be seen as a tool to enhance modelling. In this sense, I scrutinise two examples of formal modelling in IR, and derive lessons for the empirical testing of models in the discipline.

Keywords: formal models, rational choice theory, empirical testing, explanation

Resumo

Os modelos formais constituem uma parte essencial da ciência política contemporânea e das Relações Internacionais. Sua história recente está fortemente ligada aos desenvolvimentos da teoria da escolha racional, que é considerada a única teoria dedutiva nas ciências sociais. Este caráter único, especialmente sua manifestação por meio de simbolismos matemáticos, causou profundas divisões e críticas na disciplina. Os modelos formais têm sido constantemente acusados de serem construídos com base em suposições irrealistas do comportamento humano e da estrutura social, resultando em previsões triviais ou nenhuma previsão empírica. No entanto, essas críticas frequentemente ignoram elementos essenciais do conceito de explicação e como o mesmo é aplicável à modelagem formal. Neste artigo, forneço uma abordagem à modelagem matemática que considera os desafios de conceber e executar testes empíricos de previsões geradas por modelos formais. Em vez de desqualificar ou falsificar modelos, os testes empíricos são fundamentais para a adaptação de explicações mais fundamentadas dos fenômenos políticos e devem ser vistos como uma ferramenta para aprimorar a modelagem. Nesse sentido, analiso dois exemplos de modelagem formal em RI e extraio lições para o teste empírico de modelos na disciplina.

Palavras-chave: modelos formais, teoria da escolha racional, teste empírico, explicação

Introduction

Mathematics has aided a variety of sciences since the dawn of times. Ancient civilisations relied on mathematical concepts and models to describe the world around them. Models in particular constitute the essence of modern physics and chemistry, but they are also important tools in disciplines such as biology and social sciences. Political science and International Relations (henceforth, IR) have benefited extensively from models. Models of political phenomena have been designed mostly under the framework of rational choice theory (henceforth RCT), which seems plausible for RCT is the only deductive theory in the social sciences. Mathematical models are intrinsically dependant on deduction to connect assumptions, enhance logical arguments, and generate predictions. It is only natural that a deductive theory would produce such kind of models.

The first efforts in modelling political phenomena may be traced back to Borda's and Condocert's paradoxes, or, more recently, to spatial models developed in the first half of the 20th century by Harold Hotelling (1929), followed by Duncan Black (1958) and Anthony Downs (1957). Nevertheless, the use of models as methodological tools is usually attributed to Kenneth Arrow's impossibility theorem, which relied on mathematical assumptions to advance arguments about preference aggregation. Game theory also became popular in political science and IR around the same time, and together with spatial models, they account for the bulk of mathematical modelling in the discipline.

Since the initial developments, mathematical models have caused profound disputes within the discipline. Donald Green and Ian Shapiro's (1994) classical critique – *Pathologies of Rational Choice Theory* – summarises a great deal of the arguments against RC models, which tended to be echoed by many political scientists despite the responses given by RC theorists. In their piece, Green and Shapiro are profoundly concerned about the great importance and visibility given to RCT and formal models, mentioning the increasing share of RC articles being published in the *American Political Science Review.* Despite acknowledging the potential of RCT, they believe the theoretical enterprise failed in its mission of providing explanations and predictions of concrete political phenomena, such as voter turnout, legislative behaviour, electoral competition, and collective action. Their focus is eminently on the empirical power of RCT, which they deem limited.

The debate over the prospects of empirically testing RC models has echoed in the discipline, and has been constantly used as an argument against formal modelling. If models cannot generate predictions that are true to the real world, why should we bother designing such models in the first place? Wouldn't political science and IR do much better with statistical tests and/or qualitative analyses? The first question is an issue of contention not only in political science and IR, but also in philosophy. Models represent the real world, but to what extent their assumptions should be true to the world is a matter that causes profound disagreements. The second results from the historical and institutional development of the discipline. Quantitative and qualitative methods have granted political science and IR their scientific status, and still nowadays they are the main terms under which most political scientists think about methodology and, more importantly, about explanation.

However pertinent these questions may be, they tell only part of the story of formal modelling. Models come in different flavours, and distinguishing between them is paramount to understand their *raison d'être*, the predictions they generate, and the prospects of testing. Not all models are empirically testable via statistics or qualitative methods. Some models are fashioned to advance concepts, or to explain regularities observed in a specific set of phenomena, while others generate predictions that may be tested via statistical models. The main issue here is that each type of model serves specific purposes, and empirically testable models are just one type of models. Even when the prospects of empirical testing are possible, methodological issues regarding mathematical structural compatibility and measurement might be of extreme importance to define whether an empirical test is *de facto* "testing" the predictions of a given model. Throughout the remainder of this paper, I shall address the challenges of formal modelling and empirical testing. The paper is divided into three sections: the next section discusses philosophical aspects of models as representational devices, and the kinds of explanation they produce. Then, I proceed to discuss the modus operandi of empirical testing in political science and IR[2], raising questions about measurement and structural representation. I discuss these challenges in detail in the third section, where I address potential sources of problems in empirical testing of formal models, and how political scientists should cope with them in their research.

Models and Explanation

Models and modelling have been discussed in philosophy and social sciences, although each discipline focuses on different aspects of the same issue. Philosophers are usually concerned about the representational capabilities of models, i.e. how models represent the real world via mathematical assumptions. The nature of models is described in different forms: models as *"autonomous agents"* that "function as *instruments* of investigation" (MORRISON; MORGAN, 1999, p.10); models as "abstract objects constructed in conformity with appropriate general principles and specific conditions" (GIERE, 2004, p. 747); models as "experiments *in thought* about what would happen in a real experiment" (CARTWRIGHT, 2010, p. 19). Despite the semantic nuances of each definition, they share the common understanding that models are designed to represent at least

[2] A first version of this article has been debated at the 2017 Annual Conference of the Australian Society for Quantitative Political Science held in Wellington, New Zealand in December 2017.

some aspects of the world, for total representation is unattainable (and, perhaps, undesirable).

In the social sciences, the representational character of models has caused profound disputes. One of the main questions raised by sociologists, economists and political scientists, is to what extent models represent social phenomena, and how they produce explanations about the real world. If the primary goal of modelling consists in explaining the world via mathematical assumptions, then it is only reasonable to try to assess how modellers use maths to represent real-world phenomena. The literature usually focuses on two fronts to question the explanatory power displayed by models: either researchers question how realistic a model's assumptions are; or they question a model's predictions by confronting them with empirical data.

The case of assumptions is paramount to the understanding of modelling, for it raises questions about the process of designing a model. Many assumptions usually entailed in models are known to be false: perfect information, transitivity, *Homo economicus,* just to name a few. For those who adjudicate the value of models based solely on their assumptions, falsehoods cannot be ignored to assess a model's explanatory power. According to them (CARTWRIGHT, 2010; REISS, 2013), unrealistic assumptions render the model unrealistic, even if it generates predictions via a logical process. Taking this stance, however, seems too radical, for it ignores that models are not supposed to truly represent every single aspect of reality. It is the modeller's job to eschew falsehoods when deriving explanations. As Hausman (2013, p. 252) states:

> What one needs to inspect is not the model but the application of a model in a particular explanation. Such applications typically do not make use of all the assumptions within the model and so obviously do not rely on those assumptions that they do not make use of.

It is also important to note that models are designed for particular target systems, providing explanations based on the logical implications of the initial assumptions. In this sense, a model is "partially isomorphic to the real world" to the extent that

> some assumptions that define the model match some of the assumptions met in the real world, the target system. What we need in terms of truth values to make this happen is a claim indicating precisely which aspects of the model identify which aspects of the target system" (ROL, 2013, p. 246).

Therefore, the existence of falsehoods should not automatically doom a model as unrealistic, unsuccessful or false (MÄKI, 2013). As Day (1990, p. 286) asserts, models consist of "a structuring of the situation (actual or hypothetical) so that a theory can be applied".[3] This structuring establishes the links between assumptions, and hence the explanatory mechanisms of a given set of phenomena.

The essence of modelling lies on producing explanations about mechanisms operating in real-world phenomena. To be sure, the scientific endeavour consists in predicting events, and models are tailored to generate such predictions, and hence explanations. The nature of explanations produced by models falls into the domain of what Dowding (2016, p. 2-50) defines as type and token explanation. Type explanations concern

[3] Day (1990) is mostly concerned about the links between model design and theories. In his view, "[c]onstructing a model (i.e., a structuring of a situation for theory application) can be intimately related to the level of representation of the associated theory" (DAY, 1990, p. 290). In dealing with the no-slip boundary condition in fluid mechanics, he illustrates not only the usefulness, but also the tension caused by rival theories, and how models can be used to solve for this tension.

general phenomena, such as the tragedy of commons, veto power, paradoxes of social choice, etc. Token explanations are provided in terms of cases and data, meaning that they are case-specific. Models are closely related to type explanations, for they articulate assumptions that unravel the mechanisms operating in the macro-level. Take, for example, the case of Navier-Stokes equations, a set of non-linear differential equations that model fluid behaviour: they can be used (and simplified) to predict a variety of fluid behaviour, such as water flow in a pipe, the air currents on an airplane wing, the weather, blood flow etc. The mathematical model comprises the mechanisms operating in fluid flow (viscosity, pressure, stress, external forces), and solving for the equations generate predictions about how the fluid will behave given initial and boundary conditions.

This distinction between type and token explanation does not prevent critics of modelling from attacking models based on their lack of empirical evidence. It is only natural that the term prediction generates expectations about confronting a model's conclusions/hypotheses with empirical data. Nevertheless, this interpretation is misleading, for it only tells one part of the story about models. As mentioned previously, models come in different flavours, and some types are not suited for empirical testing. I hereby identify three classes of models: conceptual, quasi-conceptual, and extrapolative.

Conceptual models aim to enhance certain conceptual and theoretical arguments, resorting to the logical language of mathematics. Set theory and game theory are the most common mathematical tools in this class of models. This is the case of Arrow's theorem, which is a set of logical deductions based on set theory. The Shapley-Shubik index and Thomas Schelling's model of segregation are also examples of conceptual models that resort to similar tools. In terms of generating explanation, conceptual models offer predictions by unravelling, through mathematical expression, the mechanisms underlying political phenomena. They are not testable in the sense that one could fit data into them (which is precisely the case of the aforementioned examples). An analogy could be drawn with the third law of thermodynamics, which states that, at absolute zero, the entropy of a perfect crystal is equal to zero. However, as cooling to absolute zero is unattainable, the third law remains as a conceptual model about what would happen if we could reach such a temperature.

Quasi-conceptual models are designed to explain regularities and patterns observed in data, but which lack an explanatory mechanism. In physics and mathematics, this is comparable to conservation laws: an explanatory model for why some physical quantities were conserved was absent until Emmy Noether published her theorems in 1915 and 1918 (BAYER, 1999). Thanks to her model, a pervasive regularity – conservation of momentum and energy – was fully explained. Similarly, political scientists might be interested in designing mathematical models capable of binding data together via an explanatory mechanism. Anna Bassi's (2013) model of endogenous government formation and Torun Dewan and Arthur Spirling's (2011) model of collective decisions in Westminster systems are examples of the quasi-conceptual class. It is worth noting that quasi-conceptual models might resort to additional instruments to advance their claims, such as visual representations and/or historical evidence, as in Giannetti; Sened (2004). Nonetheless, the purpose of such quasi-conceptual models is still to offer predictions of certain phenomena via a mathematical construction that captures the regularities in data. They are, therefore, accommodationist.

Finally, extrapolative models comprise all models that are suitable for an empirical test. Nevertheless, tests might differ in respect to how they are performed. The standard approach is to test the propositions and theorems of a model by the means of a statistical model. This outcome-oriented test validates explanations based upon the results obtained through statistics. Peter Partell and Glenn Palmer's (1999) test of James Fearon's (1999) audience costs model falls into this category, as well as many others. I call this approach the data-fit extrapolative model. Yet there is still the possibility of deriving the statistical test directly from the formal model, attempting to represent the particularities of the mathematics into the statistics. Curtis Signorino (1999, 2003) offers the best example of this approach: his strategic interaction game derives a statistical test that accounts for the uncertainties entailed in the formal model. Clifford Carrubba et al. (2007) follow a similar strategy, but instead of designing a stochastic model, they derive a comparative statics model of decision-making. I call this approach a mathematical-statistical (or math-stats) model, whose main characteristic consists in bridging the mathematical model and the statistical test. They generate explanation by representing the mechanisms entailed in the model's assumptions in a statistical test. Table 1 sums up these classes.

Extrapolative models are intimately related to the criticisms found in the literature of political science. Donald Green and Ian Shapiro's famous critique scrutinised a set of popular models in the discipline, using empirical success as the main criterion for evaluating models' explanatory power (GREEN; SHAPIRO, 1994). The responses were immediate, with many political scientists criticising their biased approach to model testing (FIORINA, 1995; LOHMANN, 1995; COX, 1999, 2004). Nevertheless, the empirical criticism has echoed in the discipline, and some have come to defend that models should be rather seem as fables, whose conclusions and predictions are better understood as the moral lessons in Aesopian stories (CARTWRIGHT, 2010; RUBINSTEIN, 2012; JOHNSON, 2017). Nonetheless, empirical testing still remains an issue for modelling, at least for one type of models (namely, extrapolative). I shall deal with the prospects of empirical testing and the difficulties they entail in the next section, highlighting mathematical aspects of this debate and their implications to modelling in political science and IR.

Empirical tests and Extrapolative models

Traditionally, political scientists of the quantitative tradition have resorted to statistics to test potential predictions, and hence explanations, of political phenomena. Quantitative scholars are naturally familiar with numerical expressions that "produce a more precise description of concepts and relationships than ordinary language" (LUPIA; ALTER, 2014,

p. 54). Explanation in the quantitative tradition comes in the form of inferential propositions, tested via statistical models for causal relationships.

Table 1- Typology of models

Type		Description	Explanation	Examples
Conceptual		These models advance concepts and predictions via mathematical expressions derived from set theory and game theory. They are not empirically testable.	The logical-mathematical expressions generate explanation, unravelling the mechanisms implied by the model's assumptions.	• Kenneth Arrow's (1953) impossibility theorem • Shapley-Shubik (1988) index • Third law of thermodynamics
Quasi-conceptual		The model explains an observed empirical regularity by resorting to mathematical deductions. Data come first and the model explains their patterns.	Mathematics binds together the patterns in data, building an explanatory mechanism via the assumptions in the model.	• Anna Bassi's (2013) model of endogenous government formation • Torun Dewan and Arthur Spirling's (2011) model of collective decisions in Westminster systems • Emmy Noether's theorem
Extrapolative	Data-fit	Mathematical model and statistical test are not structurally linked via mathematical expressions. Hypotheses are formulated based on the model's propositions and theorems, and then subject to an appropriate statistical test.	Explanation results from the test of hypotheses on the level of outcomes.	• Peter Partell and Glenn Palmer's (1999) test of the audience costs model • Craig Volden and Clifforf Carrubba (2004)
	Math-stat	Statistical tests are derived directly from the mathematical model. In this case, the test represents the details of the model. There is a structural, mathematical link between the formal model and the statistical test.	Explanation results from the test of hypotheses on the level of mechanisms.	• Stephen Ansolabehere et al. (2005) measurement of voting weights • Clifford Carruba et al. (2007) comparative statics model of strategic decision-making • Curtis Signorino's (1999, 2003) strategic interaction game • LIGO experiment on gravitational waves

Source: Author's research, 2018

Most courses on quantitative methodology present the standard research cycle, which resembles the classical scientific method of the natural sciences. The researcher departs from a given phenomenon in the real world which demands some sort of explanation. She then proceeds to formulate hypotheses that shall be tested for their explanatory power. Variables and relationships between them are assigned to each hypothesis the researcher is willing to test. Results are expressed in probabilistic terms and intervals of significance, and might accept or reject the initial hypothesis (or certain hypotheses). This cycle has become so pervasive in the discipline that most quantitative articles follow a standardised structure, where they present the hypotheses, dataset, statistical test to be performed, and results.

For some time, and more specifically under the influence of King et al. (1994), quantitative scholars have believed that the logic of inference of their tradition was the only one capable of pinning down causation (DOWDING, 2016, p. 162). Evidently, that caused profound disagreements with other scholars, namely qualitative/interpretive researchers and experimentalists. Despite recent developments of more sophisticated statistical tools to analyse political phenomena, and the dialogue with experimental political science, there are still questions about the type of explanation generated by statistical tests. To be sure, quantitative methods are the only ones capable of identifying empirical generalizations, but this is not an automatic result from any research design. Moreover, identifying patterns in data is just one step in the long process of understanding the explanatory mechanisms underlying a given phenomenon.

Given the challenge posed by the essence of explanation, it would seem only natural that modellers and quantitative scholars would work closely together in a cooperative fashion. This is not always the case, and much to the contrary, modellers and quantitative researchers often see each other with some level of suspicion. R. Luce (*apud* TAAGEPERA, 2008, p. 5), speaking from modellers' perspective, summarises this tension as follows:

> Model builders find inferential statistics of remarkably limited value. In part, this is because the statistics for most models have not been worked out; to do so is usually hard work, and by the time it might be completed, interest in the model is likely to have vanished. A second reason is that often model builders are trying to select between models or classes of models, and they much prefer to ascertain where they differ maximally and to exploit this experimentally. This is not easy to do, but when done it is usually far more convincing than a fancy statistical test.

Luce raises questions about the prospects of empirically testing models that have been tackled only recently. Most of the literature in the quantitative tradition rests on the premise that hypotheses can be deduced from real-world phenomena and tested via statistical models. There is an understanding that a well-performed test suffices to at least identify potential causal relations between variables. This might hold as long as the researcher can prove that the test is statistically suitable to deal with her research design. However, when it comes to formal models, problems might occur in the process of translating a model's assumptions into a statistical test. In order to make this argument clearer, I will briefly present two approaches to testing formal models that demonstrate the translation challenge.

James Fearon's model of audience costs

In 1994, James Fearon published in APSR a formal model of international crisis, where domestic aspects influence the final outcome of crisis bargaining. In his own words (FEARON 1994, p. 577):

> I characterize crises as political contests with two defining features. First, at each moment a state can choose to attack, back down, or escalate the crisis further. Second, if a state backs down, its leaders suffer *audience costs* that increase as the crisis escalates. These costs arise from the action of domestic audiences concerned with whether the leadership is successful or unsuccessful at foreign policy.

Without going deep into details, Fearon's model rests on a game-theoretical approach, where audience costs are modelled as a linear increasing function, and the crisis unfolds as a series of stages prior to the commencement of war. Once the time horizon is reached, war is waged.

The literature that followed Fearon's game focused on testing a variety of hypotheses derived from the outcomes of the original model. Joe Eyerman and Robert A. Hart (1996), for example, designed a Poisson test based on the SHERFACS phase-disaggregated conflict management dataset, and used measures of democracy to assess audience costs. Peter Partell ;Glenn Palmer (1999) followed a similar approach, but opted for logit and maximum likelihood tests, and used institutional constraints as a proxy to measure audience costs. In both works, the assumption of the existence of audience costs is taken for granted, and no attempt is made to measure them directly. The first study where audience costs were experimentally measured was conducted by Michael Tomz (2007), which was later praised by Erik Gartzke; Yonatan Lupu (2012) as a valuable empirical proof of the validity of Fearon's model.

The puzzling issue about the literature on audience costs – and which echoes in other fields – is that data-fit extrapolative models are tested solely for their outcomes. Mathematically speaking, there is no structural correspondence between the test and the model. In the aforementioned cases, authors assumed that audience costs existed, and attempted to check whether Fearon's conclusions would hold when confronted with empirical data. Nevertheless, by focusing in the outcome, they ignored a more basic question about the underlying assumption in the model, i.e., the very existence of audience costs. Luckily, Tomz's research ended up proving that they do exist, but his results are rather silent about the mathematical behaviour of the audience costs function. In the original model, it is assumed to be linear for non-stated reasons (probably, for simplicity, but Fearon says nothing about that). Counterfactually, one could raise the question that the function is non-linear, which could potentially lead to different conclusions. Therefore, this casts doubts whether the model was effectively tested, for an essential feature in its construction was assumed to be true instead of being subject to scrutiny.

This pattern is observed in other fields where extrapolative models are subject to data-fitting. Stephen Ansolabehere et al. (2005), writing about coalition theory, criticises the use of number of seats as a proxy to measure voting weights, which are a constitutive element of a variety of coalition models. At a first glance, it would sound simply as a technicality confined to the realms of variable measurement. However, it has profound implications to the prospects of testing a model in terms of what is being represented in the test. If the test fails to translate the constitutive parts of the model, how can one be sure about its validity? From a mathematical perspective, one cannot be sure about the results of a test that do not truly reproduce a model's assumptions and propositions. Evidently, and as pointed by Luce (2008), translating a model is no easy task, but some political scientists have already started to work on math-stats extrapolative models, as I will present next.

Curtis Signorino's strategic interaction game

In 1999, Curtis Signorino published in APSR the first of a series of articles on a substantively different approach to extrapolative models. Drawing on the work of Bruce Bueno de Mesquita and David Lalman (1992), Signorino developed a mathematical-statistical model which was directly derived from the strategic game of international conflict. He made his proposal even clearer in a paper published in *Political Analysis* in 2003, where he derives three models based on distinct sources of error in the analysis of strategic games.

The quintessence of Signorino's proposal consists in translating models in ways that respect their structure, especially if they contain nonlinearities and sources of error. Paying attention to the structure of the game requires designing mathematical bridges between the assumptions/propositions and the statistical test before feeding the model with data. Signorino does so by specifying utility functions with regressors and error terms, and often by testing them via Monte Carlo simulations.[4] Clifford Carrubba et al. (2007a, 2007b) acknowledge the need for building the structural connection between model and test, but disagree with Signorino about his stochastic approach, favouring instead models that rely on simpler maths and comparative statics. Nonetheless, what we can learn from this disagreement is that different statistical models can be implemented to test a formal model, as long as they are properly derived from the latter.

The maths-stats class of models offers an extra advantage in setting the boundaries for a model's application, and hence testing. Boundaries define the range of applicability of a given model based on its assumptions. For example, in structural engineering the widely used infinitesimal stress-strain tensor offers simpler expressions to tackle problems of small displacements, where the effects caused by a given load do not interact (allowing, consequently, for the application of the linearity principle of superposition). Larger displacements require other approaches, such as the Lagrangian stress-strain tensor. The latter encompasses a larger set of cases, including those ones solvable by the infinitesimal tensor. Yet if one is working within the domain of small displacements, then the infinitesimal tensor is a natural and practical choice. Nevertheless, it would not generate right predictions if applied in contexts that do not respect its boundaries. In political science and IR, parallels can be drawn in the same lines. Failing to represent a model accordingly to its structure may result in inaccurate predictions. In his response to Carrubba et al. (2007a) and Signorino (2007) presents an argument that follows this line and summarises his approach to math-stats extrapolative models:

> Although deterministic models may under certain conditions approximate the relationships in models with uncertainty, in many other situations the predictions will be very different. If one's theoretical model includes uncertainty (e.g., private information or agent error), then the equilibrium conditions should be derived based on the assumed uncertainty. That was actually one of the points of Signorino (2003). If one wants to conduct comparative statics analysis, one should then do so based on the equilibrium conditions for the theoretical model with uncertainty. Similarly, derivation of an estimator, observable implications, or insights for model specification should be based on the equilibrium conditions of the model with uncertainty. (SIGNORINO, 2004 , p. 494)

[4] Monte Carlo methods consist of computational algorithms based on randomness used to solve mathematical problems where repeated iterations are necessary. Randomness is introduced artificially and is typically used for: sampling, estimation, and optimisation (KROESE et al. 2014). Monte Carlo simulations allow for "exploring and understanding the behaviour of random systems and data" by carrying out "random experiments on a computer and [observing] the outcomes of these experiments" (KROESE et al. 2014, p. 387).

Summing up, Signorino's works address the structural part of modelling and testing. By raising the problem of misspecification (SIGNORINO; YILMAZ, 2003) and the issue of nonlinearities (SIGNORINO, 1999, 2003), he and his colleagues call for a different methodological approach to extrapolative models. Building the bridge between the assumptions and propositions entailed in the model, and the equations in the statistical test adds mathematical consistency to the analysis.

Evidently, the literature comprises models that fall into the two classes of extrapolative models. Despite the acknowledging of the need for deeper reflection about how statistical tests translate model's assumptions, many researchers still find it hard to bridge both models. This is why data-fit models are still popular and a recurring tool for empirical testing of formal models. This is not necessarily undesirable, as long as researchers keep in mind the explanatory limits imposed by their tests. Setting the boundaries of explanation is a legitimate issue, and it might prevent political scientists from reaching conclusions that are not as comprehensive as they would want them to look like.

The pitfalls of empirical testing

The examples of the previous section draw attention to two sets of pitfalls of empirical testing, namely: 1. the challenge of measuring variables and assumptions of formal models; and 2. the structural correspondence between mathematical model and statistical test. The first is a major concern for empirical works of any kind, but it plays an important role in assessing the explanatory power of extrapolative models. The second, however, is specific to the math-stats class of models, more importantly, to the issue of structural translation.

Measurement

The literature on audience costs illustrates one of the most pressing challenges in quantitative research and formal modelling: that of measurement. Often political scientists have to measure unobservable variables (such as values and attitudes) in their models, which requires a great deal of methodological effort to capture valuable and useful information. Measuring is complicated in all sciences, and, perhaps, one of the toughest tasks a scientist might perform in her daily routine. This is so, because one has to offer concrete evidence about the variable being measured. Furthermore, in order for a set of measurements to be valid, coherent theoretical models that connect empirical evidence with properties of a given phenomenon must be precisely defined. In structural engineering, for example, displacement is the measure which is conceptually connected to the stress-strain models of mechanical behaviour: once one measures displacement, all calculations can be performed to solve for stresses, strains, and mechanical properties.

Measuring demands caution and, at the same time, creativity when testing formal models. The recent observation of gravitational waves illustrates the importance of both skills. Albert Einstein's theory of relativity predicted gravitational waves, but for a long time, scientists were unable to find any evidence of their existence. It was only in 2016 that the phenomenon was observed for the first time by the *Laser Interferometer Gravitational-Wave Observatory* (LIGO) and *Virgo*. In order to measure gravitational waves, scientists departed from the assumption that they should behave as waves, and as such they would be expected to distort the fabric of space-time. Building upon the concept of classical interferometers, LIGO was constructed as a larger version of such devices to detect specifically that distortion. Scientists were cautious enough to guarantee that their measurements were firmly grounded in the model, which was essential to the success of their endeavour.

Political scientists need similar skills to ensure that their measurements fit models' assumptions and outcomes. As we have seen, the classical approach to measurement follows the dicta of statistical tests. This is not a problem *per se*, but formal models tend to be silent about how tests should be performed and, more importantly, how variables should be measured. Correct measurements – in terms of *what* is being measured – are paramount to assessing the robustness and results of a given test, especially when dealing with math-stats models. One cannot guarantee whether a test corroborates or not a model's predictions if measurements do not correspond to the assumptions. Nor a test can be said to be appropriate if data are pushed into the model without taking into consideration measurement reliability.

Quantitative political scientists have been developing a variety of tools to ensure measurement validity and reliability (JACKMAN, 2008). Attention has also been drawn to the effects of errors and bias, not to mention to uncertainty (SIGNORINO, 2003). Formal models, and namely math-stats models, may benefit from the incorporation of such tools and concerns into their structure. In terms of data-fit models, building a compelling case is of uttermost importance to guarantee that all essential parts of the statistical test are properly connected to model's assumptions and outcomes.

Nonlinearities and structural translation

Despite the temptation of trusting linearity, real-world phenomena are pervaded with nonlinear effects. Our brains are wired to think in terms of linear relationships, preferring models where variables behave in a more-or-less linear fashion to those where variables take nonlinear paths. However, nature and society display a variety of phenomena that do not follow the tenets of linearity. Turbulence, fracture propagation, combustion, conflict escalation are just a few examples of nonlinearities.

When modelling, political scientists sometimes have to represent nonlinearities. This is the case of uncertainty, which was the main issue in Signorino's works. Failing to incorporate nonlinearities in a model might affect its explanatory power, especially if it is sufficiently complex (for example, when it entails subgames, institutional constraints, signals). However, the challenge becomes more prominent when nonlinear models are subject to statistical tests. In this case, structure matters, and failing to adequately represent a model's structure may generate incorrect outcomes or, in a less dramatic scenario, outcomes that are confined to certain boundaries where linearity applies.

Representing nonlinearities in math-stats models in particular poses the challenge of mathematical tractability. Equations may become too complex to be solved in due time, for the more complex they are, the more computational power is required to perform calculations. Political scientists have to be aware about this issue whilst translating their models into statistical tests. Unfortunately, there is no definitive recommendation, besides the fact that one should take into account some factors such as: 1. analytical tractability of the equations; 2. computational power to solve for them; 3. boundaries between which the solutions are applicable. In the case of game-theoretical models that are constructed in sequential moves, Signorino's works may serve as a point of departure to implement a math-stats model. For larger sequential games, where the game tree leads to a great number of nodes and outcomes, breaking them into smaller parts and solving for them separately may offer partial results whilst eschewing the divergence caused in the process of solving for nonlinear equations.

Another approach with which political scientists are less familiar is suggested by Rein Taagepera (2008, chapter 4). When discussing the functional forms quantitatively predictive logical models can assume, he recommends political scientists to extrapolate from the classical linear regression equation and look for functionals based on boundary conditions and logical considerations. In mathematical terms, it means expressing the problem in terms of differential equations. Solving for differential equations results in functional forms that respect boundary and initial conditions, and the resulting functional is often nonlinear. Nevertheless, basic mathematics does not suffice to deal with such equations, meaning that specific training would be necessary to model political phenomena in these terms. The consequences, however, would be profoundly positive to the understanding of politics, for political scientists would first think about each situation in a phenomenon-oriented fashion instead of automatically fitting whatever they observe in the real world into a statistical test.

Summing up, measurement and nonlinearities represent part of the challenges faced by political scientists when testing formal models. They pose difficulties to claims over a model's explanatory capabilities, for poorly designed tests do not shed light on the crucial issue of whether a model is capable of predicting real world phenomena. A well-designed test has to reflect to some extent the assumptions entailed in the model in order to have a say about its explanatory power. As Dowding (2016, p. 173) states:

> One must always ask when reading formal models how robust the conclusions are with regard to the assumptions. The more robust the conclusions, the less important specific assumptions; if a result rests upon a key assumption, how far that assumption is descriptively accurate (either motivationally for an agent or structurally for the system) will determine how useful the model is.

Final remarks

Throughout this paper[5], I have raised questions about the prospects of empirically testing formal models. Despite the widespread temptation to subject models to statistical scrutiny, only a certain class of models – extrapolative – are suitable for this kind of test. Conceptual and quasi-conceptual models, which I have not explored in detail here, present distinct features that are not fit to statistical tests.

When performing tests of formal models, researchers must be attentive to measurement issues. Models often express their propositions in terms of variables that have not been previously measured or whose data may not be directly available. Much on the contrary, modellers are rather silent about how their models can be implemented via a statistical test. Creativity and caution are necessary to ensure a test matches with a model's assumptions and propositions.

Ideally, researchers would attempt to represent the mathematics of a formal model into a viable statistical test. This is no easy task, for the translation of a model's structure might entail nonlinearities and complicated behaviour that would render the test intractable. Nonetheless, adequate translation is paramount to shed light on models' explanatory power. It is also the key to test rival models against evidence, and assess which is descriptively more accurate.

Numerics and computational simulations are useful tools to deal with nonlinearities and to test math-stats extrapolative models. Political scientists should take advantage of them to assess the tractability of their models and solve for them whenever analytical solutions are not available. Furthermore, these tools allow for the implementation of nonlinear elements, such as imperfect information and heuristics, and for iterative solutions. They might be part of the creative solutions researchers need devise to offer accurate explanations about political phenomena. This is a challenging process, though an essential one in order to unravel the mechanisms operating in the real world.

[5] I thank Dr. Shawn Treier (Australian National University) and Shirley Duarte (Federal University of Sergipe) for their valuable comments on the first version of this article.

References

ANSOLABEHERE, S. *et al.* Ting. Voting Weights and Formateur Advantages in the Formation of Coalition Governments. American **Journal of Political Science**, vol. 49, n. 3, 2005, p. 550-563.

ARROW, K. **Social Choice and Individual Values**. New Haven: Yale University Press, 1953.

BASSI, A. A Model of Endogenous Government Formation. **American Journal of Political Science**, vol. 5, n.4, 2013, p. 777-793.

BAYER, N. The Heritage of Emmy Noether in Algebra, Geometry, and Physics. **Israel Mathematical Conference Proceeeding**, vol 12. 1999. Disponível em: <http://cwp.library.ucla.edu/articles/noether.asg/noether.html>. Acesso em 4 mar. 2018.

BLACK, D. **The Theory of Committees and Elections.** Cambridge: Cambridge University Press, 1958.

BUENO DE MESQUITA, B.; LALMAN, D. **War and Reason:** Domestic and International Imperatives. New Haven: Yale University Press, 1992.

CARRUBBA, C. J.; YUEN, A.; ZORN, C. In Defense of Comparative Statics: Specifying Empirical Tests of Models of Strategic Interaction. **Political Analysis**, vol. 15, n. 4, 2007a, p. 465-482.

CARRUBBA, C. J.; YUEN, A.; ZORN, C. Reply to Signorino. **Political Analysis**, vol. 15, n. 4, , 2007b, p. 502-504.

CARTWRIGHT, N. Models: Parables v Fables. In: FRIGG, R.; HUNTER, M. **Beyond Mimesis and Convention: Representation in Art and Science**. Amsterdam: Springer Netherlands, 2010

COX, G. The Empirical Content of Rational Choice Theory: A Reply to Green and Shapiro. **Journal of Theoretical Politics**, vol. 11, n. 2, 1999, p. 147-169.

COX, G. Lies, Damned Lies, and Rational Choice Analyses. In: SHAPIRO, I; SMITH; R. M.; MASOUD, T. E. **Problems and Methods in the Study of Politics**. Cambridge: Cambridge University Press, 2004.

DAY, M. A. The No-Slip Condition in Fluid Dynamics. **Erkenntnis**, vol. 33, n. 3, 1990, p. 285-296.

DOWDING, K. **The Philosophy and Methods of Political Science**. London: Palgrave Macmillan, 2016.

DOWNS, A. **An Economic Theory of Democracy**. New York: Harper, 1957.

EYERMAN, J.; HART, R. A. Jr. An Empirical Test of the Audience Cost Proposition. **Journal of Conflict Resolution**, vol. 40, n. 4, p. 597-616, 1996.

FEARON, J. D. Domestic Political Audiences and the Escalation of International Disputes. **American Political Science Review**, vol. 88, n. 3, 1994, p. 577-592.

FIORINA, M. Rational Choice, Empirical Contributions, and the Scientific Enterprise. **Critical Review: A Journal of Politics and Society**, vol 9, n. 1-2, 1995 , p. 85-94.

GARTZKE, E.; LUPU, Y. **Still Looking for Audience Costs. Security Studies**, vol. 21, n. 3, 2012, p. 391-397.

GIANNETTI, D.; SENED, I. Party Competition and Coalition Formation. **Journal of Theoretical Politics**, vol. 16, n. 4, 2004, p. 483-515.

GIERE, R. How models are used to represent reality. **Philosophy of Science**, vol. 71, n. 5, 2004, p.742-752.

GREEN, D.; SHAPIRO, I**. Pathologies of Rational Choice Theory**. New Haven: Yale University Press, 1994.

HAUSMAN, D. M. Testing' Game Theory. **Journal of Economic Methodology**, vol. 12, n.2, 2005, p. 211-223.

HOTELLING, H. Stability in Competition. Economic Journal, vol. 39, n. 153, 1929 , p. 41-57.

JACKMAN, S. Measurement. In: BOX-STEFFENSMEIERS, J.; BRADY; H. E.; COLLIER, D. **The Oxford Handbook of Political Methodology**. Oxford: Oxford University Press, 2008.

JOHNSON, J. Models-As-Fables: An Alternative to the Standard Rationale for Using Formal Models in Political Science. **Annual Meetings of the Midwest Political Science Association, Roundtable: New Directions in Formal Theory**, mar. 2017.

KING, G.; KEOHANE, R. O.; VERBA, S. **Designing Social Enquiry**. Princeton: Princeton University Press, 1994.

KROESE, D. P. *et al.* Why the Monte Carlo Method is so Important Today. Computational Statistics, vol. 6, n. 6,, 2014, p. 386-392.

LIMA, Enzo Lenine Nunes Batista Oliveira. **Mathematics in political science: an explanation-oriented typology of rational choice models.** 2018, Tese (Doutorado) - Programa de Pós-Graduação em Ciência Política da Universidade Federal do Rio Grande do Sul, Porto Alegre.

LOHMANN, S. 1995. The Poverty of Green and Shapiro. **Critical Review: A Journal of Politics and Society**, vol. 9, n. 1-2, 1995 , p. 127-154.

LUPIA, A.; ALTER, G. Data Access and Research Transparency in the Quantitvative Tradition. **Political Science & Politics**, vol. 47, n. 1, 2014, p. 54-59.

MÄKI, U. On a Paradox of Truth, or how not to obscure the issue whether explanatory models can be true. **Journal of Economic Methodology**, vol. 20, n. 3, 2013, p. 268-279.

MORRISON, M.; MORGAN, M. Models as Mediating Instruments. In: MORRISON, M.; MORGAN, M. **Models as Mediators: Perspectives on Natural And Social Science**. Cambridge: Cambridge University Press, 1999.

PARTELL, P. J.; PALMER, G. Audience Costs and Interstate Crises: An Empirical Assessment of Fearon's Model of Dispute Outcomes. **International Studies Quarterly**, vol. 43, n. 2, 1999, p. 389-405.

REISS, J. The Explanation Paradox Redux. **Journal of Economic Methodology**, vol. 20, n. 3, 2013, p.280-292.

ROL, M. Reply to Julian Reiss. **Journal of Economic Methodology**, vol. 20, n. 3, 2013, p. 244-249.

RUBINSTEIN, A. **Economic Fables**. Cambridge: Open Book Publishers, 2012.

SHAPLEY, L. S.; SHUBIK, M. 1988. A Method for Evaluating the Distribution of Power in a Committee System. In: ROTH, A. E. **The Shapley Value: Essays in Honor of Lloyd S. Shapley.** Cambridge: Cambridge University Press 1988.

SIGNORINO, C. S. Strategic Interaction and the Statistical Analysis of International Conflict. **American Political Science Review**, vol. 93, n. 2, 1999, p. 279-297.

SIGNORINO, C. S. 2003. Structure and Uncertainty in Discrete Choice Models. **Political Analysis**, vol. 11, n. 4, 2003, p. 316-344.

SIGNORINO, C. S.; YILMAZ, K. Strategic Misspecification in Regression Models. American **Journal of Political Science**, vol. 47, n. 3, 2003, p. 551-566.

SIGNORINO, C. S. On Formal Theory and Statistical Methods: A Response to Carrubba, Yuen and Zorn. **Political Analysis**, vol. 15, n. 4, 2007, p. 483-501.

SUGDEN, R. Explanations in Search of Observations. **Biology and Philosophy**, vol. 26, n. 5, 2011, p. 717-736.

TOMZ, M. Domestic Audience Costs in International Relations: An Experimental Approach. **International Organization**, vol. 61, n. 4, 2007, p. 821-840.

VOLDEN, C.; CARRUBBA, C. J. The Formation of Oversized Coalitions in Parliamentary Democracies. **American Journal of Political Science**, vol. 48, n. 3, 2004, p. 521-537.