Ca' Foscari
University
of Venice

Single Cycle Degree programme

in Comparative
International
Relations

Final Thesis

Policy vs. Popularity:

A Case Study of Political Decision Making in the European Parliament in the Context of Meat-Substitutes Denominations

Supervisor
Ch. Prof. Valerio Dotti

Graduand
Gabriele Donato
Matriculation Number 872558

Academic Year
2023 / 2024

Ringraziamenti

Ringrazio il professor Valerio Dotti per avermi permesso di sbagliare, provare e riprovare, lavorando liberamente sui temi di questa tesi. Ringrazio il professor Roberto Casarin, che mi ha fornito gli strumenti per poter condurre il mio pensiero senza guida alcuna nei modelli più avanzati. Ringrazio ancora gli amici che mi hanno esortato ad aspirare sempre più in alto credendo in me, e la mia famiglia per le possibilità ed i mezzi che mi hanno permesso di spingere i miei studi oltre i requisiti minimi per il loro completamento. Tutti questi contributi mi hanno permesso di affrontare una trattazione formale dell' inferenza causale, portando a compimento per vie che non potevo immaginare un percorso iniziato con lo studio delle filosofie orientali, sempre attente alla produzione condizionata degli eventi ed ai rapporti di causa-effetto.

INDEX

Introduction

The present thesis investigates a legislative proposal regarding the ban of meaty names for vegan and vegetarian meat-substitutes voted by the European Parliament on 23/10/2020. The research question that will be addressed is whether, and to what extent, it is possible to characterise the behaviour of the Members of the European Parliament (MEPs) as *policy-seeking* or *career-seeking.* The thesis that will be defended is that the observed votes, retrieved from the roll-calls, are more likely to have been produced by a *career-seeking* attitude.

It is worthwhile to trace back the process that has led to the formulation of this research. In principle, it was assumed to be possible to link the observed behaviour of the MEPs to the influence of the meat lobbies. As will be explained in the chapters regarding Copa Cogeca, one of the main stakeholders in the agricultural and farming sector, the current studies regarding the Common Agricultural Policy and its supporters had inflated and distorted the expectations about both the available data and the desired results. Copa Cogeca is indeed a powerful lobby that has often influenced European policies. However, if one lesson is to be retained from the present study, is that this influence has been diluted through time, and the common assumptions of involvement in the policy process are not necessarily valid in the specific case.

The present work is partly anomalous, since its object is a proposal that was rejected. More pragmatically, it means that, if the abstentions are not counted, the variable that registers the vote is binary, and that, if the value of interest is the vote in favour of the proposal, or the "positive class" (i.e. the value of the variable that assumes value one), is doomed to be the minority class. From the very beginning, it will be possible to see that the case presents at least two challenges: (1) dealing with asymmetric outcome data, and (2) providing an inference on the hidden determinations of the MEPs having only partial measurements for the class of interest, and the possibility to base the reasoning on data that do not necessarily correspond to those determinations. To these complications, should be added those regarding the creation and retrieval of appropriate predictors.

Moreover, it should be remarked that the presence of lobby-names in the following pages does not entail that the results that will be arrived at will be linked necessarily to the action of specific interest group(s). The purpose of this thesis is to determine whether the behaviour of the MEPs may be ascribed to one of two categories, it will not be possible to determine what are the "real" incentives behind the actions. Nonetheless, it will be possible to show the non-neutrality of some

stakeholders, and the insufficiency of certain explanations instead of others. These tasks are not banal and will require advanced modelling techniques.

The present work presents non obvious results in that it counters a simplistic account of lobby influence over the policy process. Most notably, it utilises precise assumptions about the phenomena studied, more than deriving conclusions from generally accepted tenets. Some of these conclusions were unexpected, since it will be shown that some groups, such as the Greens, were indeed *policy-seekers* in the sense that the voting pattern of their members seems to be better explained by motivations not directly linked to career and vote-seeking.

Finally, the thesis has been a personal journey into worlds unimagined, both mental and physical thanks to the infinite opportunities that the discovery of the computational and programming tools is able to open.

1 The case study

1.1 The European legal framework regarding meat

The 2007 Commission's White Paper titled *A Strategy for Europe on Nutrition, Overweight and Obesity Related Health Issues,* has already recognised the impact of information over individual rational choices. Arguably, the labelling issue that will be studied in this work lies between the concerns over the internal market regulation and the enabling conditions of healthy lifestyle choices for the citizens (pp. 3-5). However, the paper was still focusing on sector specific interventions (e.g. the promotion of fruits and vegetables for children through the Common Market Organisation), or the development of partnerships with the private sector (pp. 6-10).

Regulation 1169/2011, also known as "Food Information Regulation" (or FIR), recalls the aforementioned White Paper. The Regulation had the purpose of modernising the labelling legislation by amending the other legal documents disciplining the field.[1] Provided the interest of consumers in making informed decisions (art. 1), the language used for labels must be clear and expressed in a

[1] The full title is: "amending Regulations (EC) No 1924/2006 and (EC) No 1925/2006 of the European Parliament and of the Council, and repealing Commission Directive 87/250/EEC, Council Directive 90/496/EEC, Commission Directive 1999/10/EC, Directive 2000/13/EC of the European Parliament and of the Council, Commission Directives 2002/67/EC and 2008/5/EC and Commission Regulation (EC) No 608/2004".

language familiar to final buyers (art. 15). The name of the food must be its legal name if existent, otherwise its customary name, and, if not existent, a descriptive name (art. 17).[2] Annex VI provides further guidelines in this respect, in particular the label should indicate: if a food is produced using a substitute ingredient other than the one that would be naturally expected by the consumer (point 4); if meat products and preparations giving the impression of being made out of a whole piece of meat or fish, are truly made out of a composition of pieces and cuts (point 7); if additional animal protein are added to meat and fish products (point 5); and if the meat and fish products that appear as composition of pieces contain more than 5% of added water (point 6).

Since establishing whether a name is legally protected has economic relevance, these are not only definitional questions. As regards meats, art. 1(f) of Chapter 1 of Reg. 1169/2011 recalls Annex I of EC Reg. 853/2004. According to the latter, "Meat means edible parts of the animals (…), including blood" (point 1.1). Article 1(n) of Reg. 1169/2011 defines a legal name as "the name of a food prescribed in the Union provisions applicable to it." The relevant provisions, regarding both dairy and meat products, can be found in Reg. 1308/2013. However, while the first are characterised by a comprehensive list of protected names (Annex VII, Part III, point 2(a)), the second do not share the same level of protection. Indeed "meaty names" (with the exception of "meat" itself), are not protected under EU law (Sochirca, 2018, p. 518).

The legal disparity between sectors was key in the so-called *Tofutown* case (*see* next chapter), but also object of question E-004044/2017, addressed to the Commission by EP deputies Paolo De Castro (S&D) and Giovanni La Via (PPE). They asked whether sector-harmonising measures had to be expected. The answer was rendered July 27 2017 by Commissioner Mr. Andriukaitis who explained that the extant measures where sufficient and that the Commission was not working on updating the legislation in order to introduce reserved names for meat products in Reg. 1308/2013.

---

[2] According to art. 2: "legal name" is either the name envisaged by the applicable Union provisions, or, when absent, the applicable national laws of the State where the product is sold (point n); "customary name" refers to names commonly accepted where the product is sold, so that no further qualification is needed (point o); "descriptive name" is a name that describes the product, and possibly its use, in order for consumers not to get confused by similar products (point p).

The answer refers also to "implementing acts" for vegan and vegetarian foods. As envisaged by art. 36(3)(b) of Reg. 1169/2011, regarding the voluntary informations as to the suitability of foods for a vegetarian or vegan diet, an obligation to define labelling rules was pending on the Commission. The European Vegetarian Union (EVU) also joined the discussion and proposed its own definitions, this was followed by the Commission's communication that implementing acts would be worked out starting from 2019. However, the only document that could be found in this regard on the Commission page, is the follow up by the Commission to the REFIT[3] platform opinions (European Commission, 2018). It seems that to date such implementing acts are still lacking (European Vegetarian Union, FoodDrinkEurope and EuroCommerce, 2021). Therefore, vegan and vegetarian meat substitute foods are merely subjected to the general provisions of Reg. 1169/2011, which amount to an obligation to avoid misleading labels.

1.2 The French case in a European perspective

June 29, 2022 the French Assembly adopted *Décret no 2022-947 relatif à l'utilisation de certaines dénominations employées pour désigner des denrées comportant des protéines végétales,*[4] disciplining the maximum amount of plant-based proteins that can be contained in meat preparations, as specified in the *Annexe*. The relevance of the law is primarily national. Indeed it does not apply to products "legally produced or commercialised in another Member State or Turkey, or another State participating in the European Economic Area" (art. 5, my translation), but only to foods containing plant-based proteins produced in the French territory (art. 1). These are prevented from being labelled with names of animals or pertaining to animal anatomy, legally defined meat-specific denominations, or commonly used appellations in commercial practices (art. 2).

To be sure, these ideas were already under discussion in 2018 when the *Project de loi pour l'équilibre des relations commerciales dans le secteur agricole et alimentaire et une alimentation saine et durable* was proposed (Carreño and Dolle,

[3] Acronym for "regulatory fitness and performance programme". Its aim is rendering EU legislation more supple by simplifying it. Then Commission reports yearly on the issue and compiles an "Annual Burden Survey". Further information can be found at: https://commission.europa.eu/law/law-making-process/evaluating-and-improving-existing-laws/refit-making-eu-law-simpler-less-costly-and-future-proof_en.

[4] Hereinafter, the *"Decrét."*

2018, p. 579). Nonetheless, the final text, approved after some minor modifications proposed by the Senate, does not contain any reference to the issue. Since the case that will be analysed in the present thesis has an object analogous to the *Décret,* and constituted a proposal of amendment of an existing European Regulation by a French deputy in the European Parliament, it may be useful to take into account the French case and try to connect it to a wider background.

As has been suggested, while it is arguable from the primary documents that concerns with food labels existed in the French Parliament well before 2022,[5] the impossibility to retrieve further details shall make the *Décret* the starting point of the present enquiry. However, more than explaining the forces and causal relationships behind the adoption of the law, an endeavour that falls outside the purpose of this work, the main concern here is to assess, qualitatively, whether it could be said that issues of the same or comparable order were debated at a European level at the time. If it is so, it would be reasonable to imagine that there existed a linkage between what was happening at the European level and what was happening in a particular Member State given the overarching structure of the European Union, whose admitted accomplishment has been precisely the creation of a unitary political space (Weiler, 1999, pp. 3-9).

At the supranational level, the landmark judgement in this regard has been the one of Case C-422/16 *Tofutown,*[6] in which the German company TofuTown, a producer of plant-based dairy substitutes which were marketed with designations pertaining to dairy products, has been found not compliant with Reg. 1308/2013 (EU) by the European Court of Justice. According to the Court, the use of certain designation (in this case, e.g. "milk", "cheese") is legally restricted by the Regulation and may not be used for marketing products lacking the prototypical characteristics envisaged by the law (Tofutown, para 8).

The case is notable because the Court rejected Tofutown's argument that the law was perpetuating a situation of inequality, with the dairy sector being comparatively more protected and regulated than the meat one. Indeed, "the fact that…substitutes for meat or fish are not, according to TofuTown, subject to restrictions comparable to those to which the producers of vegetarian or vegan

---

[5] Another example would be the discussions that took place on the subject already in 2018, as regards an amendment to France's Marine and Fishing Code, now implemented by the *Décret* itself (Carreño, 2022, p. 666).

[6] The case was heard in 2016 but the judgement has been uploaded on the website of the Court only in 2017.

substitutes for milk or milk products are subject, pursuant to Annex VII, Part III, to Regulation No 1308/2013, cannot be regarded as inconsistent with the principle of equal treatment" (Tofutown, para 50). This approach was backed by further concerns with the protection of consumers, not to be misled by product names, and with issues of fair competition, for which Tofutown incurred also in internal responsibility (Lähteenmäki-Uutela *et al.*, 2021, pp. 6-7).

Against the grain of this decision, it becomes clearer that what the *Decrét* disciplines is precisely what the European Regulation leaves out. As mentioned, the *Annexe* specifies the maximum amount of plant based protein tolerated in meat products (e.g. the French white sausage "coudenou" may only contain 0.1%). One notable exception is the term "burger", which has been left out from the final list but was present in the original one. This modification occurred as an outcome of the criticisms following the process of transparent legislation required for technical regulations.

According to the Technical Regulation Information System (Directive 2015/1535),[7] barriers to trade may only be justified when necessary and functional to the public interest. Therefore, the obligation of due diligence pending on Member States *ex* art.4(2) TUE requires that they notify to the Commission their draft technical regulations, allowing for a sufficient time for the evaluation (art. 5 and 16 TRIS). This entitles not only the Commission, but also the Member States to suggests amendments to be taken into account in the final formulation of the law (art. 13-14 TRIS). It should also be noted that the Directive takes into account a profile of inaction on the part of European institutions: in particular art. 17, formulated in opposite terms, holds that "Member States should refrain from adopting technical regulations once the Council has adopted a position at first reading on a Commission proposal concerning that sector."

The *Decrét* was adopted pursuant to the TRIS procedure. As can be seen from the official website, it was notified October 1 2021 and by January 3 2022 it had received comments by the Commission, as well as Sweden, Slovenia, Czechia and Portugal.[8] The draft law is described as a means of protecting consumers in their food-choices (point 9) and is deemed not liable of having a significant impact on international trades.

---

[7] Hereinafter, "TRIS".

[8] This can be checked directly on the website: https://technical-regulation-information-system.ec.europa.eu/en/notification/15553, Accessed: 23/5/2023.

Among the other commentators, the European Consumer Association BEUC is reported to have criticised the draft law as being incompatible with the "Farm to Fork Strategy" of the European Union (Carreño, 2022, p. 666). Most notably, in 2020 BEUC had already been a supporter of meaty denominations for plant-based foods, and had informed the Parliament, ahead of a vote on the subject, with statistical evidence in favour of the position that consumers are not misled by such denominations (BEUC, 2020).

In hindsight, it is clear that the *Decrét* cannot be seen as a document with a merely national dimension, even though it is formally a domestic law. On the one hand, it has been judged as not harmful for international trades, and still it had to be notified pursuant to the TRIS mechanisms. It could in no way affect producers outside of France, and still a European consumer organisation criticised it and linked its provisions to the wider European political debate. The multiplicity of relationships in which the draft law became enmeshed, demonstrate that similar themes must have been already discussed at the European level (or at least be felt) well before the events that will be investigated in this thesis. Put it differently, it is reasonable to think that concerns with Regulation 1169/2011 (see next section) might have been present in the agendas of some parties even before the *Tofutown* case.

Indeed, these intuitions seem to be confirmed by the question E-00377/2016 asked to the Commission by deputy Renate Sommer (PPE), regarding the labelling of vegan and vegetarian products. The Commission, in the person of Mr. Andriukaitis, replied that the existing legal framework was already sufficient and that vegan and vegetarian products needed not, as per law, any further qualification in order to ensure customer protection. Two years later, deputy Anja Hazekamp (GUE/NGL) specifically questioned the Commission over the *Décret* in written question E-002791. The answer, again by Mr. Andriukaitis, simply referred to the obligation not to use misleading labels (art. 7(1)(a) of Reg. 1169/2011). Since the French law had not already been notified to the Commission at the time, no answer could be specifically rendered. However, as it has been explained in this section, the law turned out to be within the limits imposed at the supranational level.

To be sure, also other Member States, such as Finland, have shown similar concerns (Läteenmäki-Uutela *et al,* 2021, p. 7), and in conclusion it is possible to affirm that the debates generated by the asymmetries in sectoral protection had long been present on the European scene.

1.3 Decision making in the European Union

The power to make a legislative proposal within the EU falls mostly to the Commision. In order to balance the democracy principle with this asymmetry between institutions,[9] it is possible for the Parliament or the Council to send non binding requests to the Commission (art. 225 and 241 TFUE). To be sure, the central role of the latter institution is justified by both its function as upholder of Union values, and by the fact that proposals are usually submitted before having taken into account stakeholders' interests. Once the proposal has been made, then the treaties provide for two procedures, one ordinary and one special (Baratta, 2022, pp. 244-251).

Since the special legislative procedure is not of interest for the present thesis, only the ordinary one will be dealt with here. It involves both the Parliament and the Council as co-legislators with equal weight, and requires three successive readings. A proposal can be approved at first reading if neither institution advances any modification. While in the Council the discussion is generally limited (the majority of the work being done by the COREPER or the Committee of Permanent Representatives),[10] in the Parliament the proposal is usually left to the relevant Committee(s).

The amendments should be accepted by the Commission, and the Parliament may be requested to vote on them in plenary session. When this phase is concluded, the proposal is sent to the Council, who adopts a position and sends it back to the Parliament. This may either approve, reject or demand amendments to the proposal. Upon rejection, the Commission may intervene in the third phase through a Conciliation Committee. The time window gets narrower at each reading: the first has not time constraints, the second should last no more than three months, and in the third the Conciliation Committee has between six and eight months to reconcile the institutions (Ibid., 2022, pp. 247-251).

---

[9] The word is here used technically and refers only to the institutions of the EU which are legally so because of their recognition in the constitutional treaties: Parliament, European Council, Council of Ministers, Commission, Court of Justice, Central Bank and Court of Auditors (Baratta, 2022, p. 130).

[10] COREPER members, known as EU ambassadors, are divided into two groups and represent the Member States. Their function is supporting the work of the Council (art. 240 TFUE). Among the areas under the competence of Coreper I, there is agriculture and fishing. For detailed information see: https://eur-lex.europa.eu/IT/legal-content/glossary/coreper.html.

1.4 The case study

The present thesis deals with the rejection by the European Parliament of the legislative proposal 2018/0218 (COD) to the effects that meaty names could not be used for vegan and vegetarian products in the EU. According to some early commentators, such as Green MEP Molly Scott Cato, it bore the interests of the meat industry (Stone, 2019). This position was explicitly defied by the Special Rapporteur Éric Andrieu, charged with overseeing the legislation (Audino, 2019).

The proposal was presented according to the tropes of clarity and good information that can be found in the regulations cited in the above section. It was elaborated by the Agricultural and Rural Developement Committee. Apart from Andrieu (S&D), also Anne Sander (EPP), Decercle Jérémy (Renew Europe), Benoît Biteau (Greens), Mara Bizzotto (Identity and Democracy), Ladislav Ilčič (ECR) and Petros Kokkalis (GUE/NGL) worked as shadow rapporteurs.[11] As it is reported, it passed with 29 favourable votes against 7 contrary votes in the Committee (Fortuna, 2019).

As a consequence, it was subjected to the ordinary legislative procedure and was part of a general reform of the Common Agricultural Policy put forward 1 June 2018 by the European Commission and adopted only in December 2021. The reform consisted of three packages: the CAP Strategic Plans Regulation, the CAP Horizontal Regulation and the Amending Regulation (European Parliament, 2021).

The debated proposal can be found in the first reading report of the procedure (A-8-2019-0198). More specifically, the object of the present work will be amendment 165, introducing a new point (31a) in Part I of Annex VII of Reg. 1308/2013. According to this modification: "The meat-related terms and names that fall under Article 17 of Regulation (EU) No 1169/2011 and that are currently used for meat and meat cuts shall be reserved exclusively for edible parts of the animals." This formulation extends the level of protection to "meat-related" names and therefore is wider in scope than art. 17 of Reg. 1169/2011 (discussed above).

Despite being only forty-four lines long, mostly recalling the existent legislation, amendment 165 has been widely debated as can be seen by searching generically for the issue. Nonetheless, also amendment 171 dealt, more indirectly, with the issue. Indeed, it specified that the protected names had to be preserved in a variety of contexts spanning from the commercial to marketing usage. Since amendment 165 has been the most debated, and it is directly linked to the meat

[11] For a full chronology see: https://oeil.secure.europarl.europa.eu/oeil/popups/ficheprocedure.do?reference=2018/0218(COD)&l=en.

sector, it will be taken as a starting point for the present inquiry. The associated roll-calls can be found, along with the votes for the remaining amendments, in the file P9_PV(2020)10-23 (p. 136), available on the website of the procedure (Note 11).

2 The theoretical framework

2.1 Inefficiencies and policy process

In the context of the study of the inefficiencies in the policy process, two main schools contend opposite views. On the one hand, the Chicago School of Political Economy (CPE) is focused on the idea that partisans are mere maximisers of their own interest. The polity reflects, at any time, not dissimilarly from stock prices, all information about the society. On the other hand, the Virginia school argues that the policy process is generally imperfect, and it has been foundational to the development of public choice theory.

A textbook example that introduces the difference, is the price support mechanism for sugar producers in the United States. In the U. S. the support program benefits 10.000 producers of sugar beets and cane at the expense of 250 million consumers. Moreover, a system of import quotas raises the sugar price above world price, while the domestic consumption has decreased in recent years due to the appearance of substitute products. The criticisms to this policy stem from the consideration that it has caused a surge of sugar prices above world prices, encouraging the import of sugar-containing products competing with domestic ones, and displacing, for this reason, several thousand jobs while costing about 3 billion dollars per year to consumers (Pasour, 1992, 155-157).

The persistence of such an harmful policy is explained according to the "diffuse cost, concentrated benefits" paradigm: the program can be seen as a subsidy to the sugar industry, more organised than consumers, and better able to affirm its interests.

However, the take of the aforementioned schools on this particular problem is different. CPE's argument rests on the idea of "comparative efficiency" : while a certain policy might be sub-optimal from a given perspective, it is assumed that, from another point of consideration, it should be regarded as optimal. In the case of the sugar problem this amounts to affirming that the dead-weight cost of the program is low when compared to the alternative programs that would have been too costly to muster public support (Becker, 1983, pp. 380-382). The Virginia

school's understanding of the problem is more focused on the asymmetry of information between the voters and the politicians, with a particular emphasis on the fact that for the voter the rent-seeking behaviour of the sugar industry has resulted in a set of regulations that increase the cost of gathering information about the policy process (Crew and Twight, 1990, p.23).

To be sure, in the sense of the CPE, "inefficiency" cannot possibly exist since the resources are assumed to be allocated, according to the axiom of rational thinking, in a way such that the increase in the utility of one decision maker would result in a diminution of the utility of another decision maker :"what is is efficient" (Pasour, 1992, pp. 156-157 and p. 159).

The present thesis builds on the framework provided by the Virginia, or "informational" school of political economy. As it will become clear in the ensuing chapters, the school offers some insights and hypotheses that are both reasonable and suitable for the case study that forms the focus of this work.

2.2 Individualism and interaction

The Virginia school of political economy is primarily linked to the works of James M. Buchanan and Gordon Tullock. As argued by the authors, the school is based on methodological individualism, or the idea that to explain social phenomena it is necessary to consider individual choices and motivations (Buchanan and Tullock, 1962, pp. 10-40). While the complete rejection of a macro level might be criticised, and it is not fully endorsed in this work, from an analytical perspective it is sufficient to imposes certain requirements of granularity in the data to be used. As it will be explained later (*see* 2.4 and 2.5), one of the major challenges of this work has been precisely the retrieval of appropriate data.

To be sure, however, making the individual the primary unit of inquiry is not necessarily a prerogative of this school. An alternative to the theories regarding the policy process, less rooted in the economic debate, is the "Social Identities in the Policy Process" (or SIPP) model. In order to better understand the framework of this work, it might be useful to compare both the similarities and differences between these two approaches, and highlight the specificities of the one chosen. In this way, it will be possible to justify precisely why and how certain data have been gathered, and why the present framework is more appealing than the available alternatives.

The SIPP model should be inscribed in the strand of the Social Identity Theory (or SIT) that appeared in the early seventies with the work of Polish social

psychologist Henri Tajfel. The novelty of his research consisted in a breakaway from earlier models of inter-personal relationships, whose dynamics took the group as the central explanatory element. In contrast, Tajfel argued that human relations had to be understood in a spectrum ranging from pure inter-personality to pure inter-group. Consequently, the SIPP model constitutes an application of these psychological theories to a political setting. It builds on three levels: micro, meso and macro.

At a micro level, it can be argued that the strength of a social identity is directly proportional to five factors: the intensity of the belonging to a group, the subjective evaluation of the group, the emotional bound with the group, the frequency of the contact with other actors sharing the same identity and how long the identity persists (Hornung, Bandelow and Vogeler, 2018, pp. 218-219).

At a meso-level, it can be argued that collective action follows group structures. For instance, frequent contacts between effective group leaders can eliminate or smooth inter-group differences. In a similar fashion, the collective action of the members of a group is more likely if their multiple identities overlap (Ibid. p. 219).

At a macro level the identities can be characterised as local, with reference to the policy process and actors; sectoral, with reference to the specialisation of the policy elites; organisational, with reference to participation to committees and parties; demographic, with reference to the provenance of the actors, their sex and their age; or informal, with reference to their participation to advocacy coalitions or teams (Ibid. pp. 220-222).

While intuitively the intersection of these different levels seems to be beneficial for a comprehensive analysis, it is argued that it might not be capable of providing significant insights in the policy process. To be sure, the SIPP model is a "socialisation" theory, however the data required to perform a meaningful analysis pursuant to the hypotheses presented in the above paragraphs, seem to be hardly available in the context of the study of a polity. Apart from the higher levels of analysis, the SIPP framework would require reconstructing the personal and ideological relations between partisans, assuming that they should be significant in explaining the legislative outcome of a specific proposal. While socialisation might be important in certain case, it is not a presumption that bears the mark of universality.

Moreover, the hypotheses of the SIPP framework pose challenging modelling obstacles to the researcher. It may be argued that neither the proxies available (the votes cast and the demographic elements about the partisans) nor the logistic or multinomial regression might be adequate for disentangling the complex nets of relationships (not granted to be linear) that form the core assumptions of the model (either in general or with reference to the available data). Indeed, these obstacles should be related to the application of a model from one field of study where it was possible to form groups "experimentally" to one where the social structure is largely (albeit not entirely) unknown. In general, considering the few and not especially satisfactory applications of this model, the thesis that the availability of data and models is still not sufficient for a fruitful application seems to be justified. [12]

As recalled above, the presence of non-legislative actors as relevant factors influencing the policy process is a landmark characteristic of the informational school of political economy. Given that the process is not perfect, the form of transfers to "special-interest groups" can be explained by the imperfect information on the part of the citizens. Politicians are assumed to adopt a form of disguised transfer to increase the cost (for citizens) of political resistance and information-gathering whenever the absence of a pressing political competition allows for it. In this sense, politicians' utility is assumed to be generally centred around the target of re-election and reputation building (Coate and Morris, 1995, pp. 1210-1214). However, it should be noted that policy and career incentives are overlapping and generally intertwined in the context of the European Parliament.

In addition, the classical criticism moved to the school, i.e. that the rejection of super-individual elements ignores the societal and cultural forces that orient political choices (see e.g. Frey, 1978), seems not be applicable to the present case since careful attention has been put in collecting the variables to be used so that they can be representative of the three levels that likely influence the legislative behaviour of an MEP (individual, national and European).

---

[12] Tajfel formed groups according to the "minimal group paradigm", where groups are formed on the basis of random criteria in order to study the behaviour of the participants excluding inter-personal favoritism (see Hornung, Bandelow and Vogeler, 2018). The cited study is also an example of the scarce results and model inadequacy that has been criticised before: the great majority of the variable used is not statistically significant, and none of them (age, gender, votes cast, and an unclear classification based on political manifestos) is liable, *per se*, of providing insights in the relations between partisans, and no justification as regards the suitability of these specific variables is given by the authors.

2.3 Lobbying the EU

Among the recent venues of research regarding lobbying activities in the European Parliament, some studies have focused on the relationship between the structure of the institution and the dynamic interaction between lobbyists and partisans. These studies have found that interest groups might engage with non-natural allies to make their positions heard within the winning coalition. These complex interactions usually involve MEPs that are either ideologically aligned or powerful, and usually take place within the larger parties (Marshall, 2015, pp. 311-312 and pp.318-322). Interestingly, although the proposal that is the object of this study is generally linked to the tropes of the parties from the right, it was put forward by Eric Andrieu, from a centre-left party (S&D).

The role of some MEPs, distinguished from their colleagues by the title of "rapporteurs" has sometimes attracted the interest of research. Committee rapporteurs are charged with the task of drafting reports that form the basis of the subsequent legislative steps, and for this reason they are generally assumed to play a pivotal role in the interaction with interest groups. While some authors focus mainly on the pursuit of specific policy goals and on partisan considerations (Mahmadou, 2002, pp. 6-8), other studies (*see* Yoshinaka *et al.,* 2010) suggest that while the appointment of rapporteurs is influenced by strategic considerations related to the pursuit of specific policy goals, technical expertise is an element not to be excluded: "On the one hand rapporteurs rapporteurs can be seen as actors that help pursue party goals [...] On the other hand, rapporteurs may be seen as facilitators intended to help achieve goals that are often technical or technocratic. Rapporteurs are the agents by which consensus builds around a point of view, possibly one privileging expertise rather than party lines" (Ibid. p. 466).

This strand of research is grounded in the understanding that the European Parliament presents multiple structural layers, the most characteristic of this institution being the committees. According to the EP Rules of Procedure, the members of the committees "shall be elected after nominations have been submitted by political groups and the non-attached Members. The Conference of Presidents shall submit proposals to Parliament. The composition of the committees shall, as far as possible, reflect the composition of Parliament" (Rule 177; EP, 2007).

In practice, the distribution of the positions happens before the plenary vote. Seats are allocated to party groups proportionally to their size in the plenary session, however the individual assignments are made according to the expectations

of the national party delegations. Nonetheless, it seems that some differences exist depending on the size and influence of the group: in larger parties, where there is an higher probability of conflict, the assignation is made within national delegation with the group leadership serving as a dispute solver (e.g. PSE); in other cases, the assignation is made by the Bureau of the group according to the wishes of the national delegations (e.g. The Liberals); finally, within some groups (e.g. the Greens), individual preferences seem to be taken more into account (Yordanova, 2009, p. 257). In general, it would seem that partisan considerations are more limited and that the technical expertise as well as the necessity of transversal collaboration are the main drivers of assignation (Ibid. pp. 275 ff).

While not directly related to party-interest groups relations, these findings are relevant for the present work since they suggest that both the formation of and the positions produced by the committees are the result of complex processes and sets of differing ideas. The presence of the technical element, and the idea that committees are not necessarily formed because of partisan considerations, impose a relativisation of their weight that will be reflected in the analyses.

The current literature on the influence of interest groups at a European level has produced contradictory results, mainly due to obstacles of three orders: defining influence, considering multiple sources of influence and measuring influence (Dür, 2008, 1220 ff). It can be seen that these problems arise when considering the overall impact of interest groups in certain policy areas (e.g. Vonk, 2022, p. 2), and are not entirely relevant in the present case where a thorough investigation on one single proposal could be made. Moreover, the thesis is focused on the behaviour of MEPs assuming that they act rationally according to a set of considerations regarding their re-election and career. In this context, given a policy, it is hoped to determine which considerations better explain the policy outcome. Conclusions regarding the influence of a certain interest group might be consequential, but neither general nor definitive, and certainly not the ultimate objective of this work.

Nonetheless, the creation of two influential umbrella organisations to foster the interests of the business sector (namely Copa-Cogeca and BusinessEurope) has been actively encouraged by the European Community (and then Union) since 1958. The growth of business representation at a European level has followed the trajectory of the history of this institution, stagnating during the first years of the European experiment and then resuming its growth starting from the eighties (Ibid., pp. 5-12).

Information about the interest groups is now available through the Transparency Register, an initiative of the European Commission to keep track of the groups that aim to take part in the policy process. One of the most prominent features of this database is that for an organisation to be able to access the Parliament, possible only through a Pass, registration is required.[13] Moreover, shadow rapporteurs, rapporteurs and committee chairman are required to publish on their personal page of the EP website the details of the meetings held with accredited interest groups (Vonk, 2022, p.12). The utility of these data is questionable, they are sparse, inconsistent and, admittedly, incomplete: reporting remains voluntary for the MEPs who do not cover one of the positions specified above, a number comprising the great majority of the Parliament.[14]

As regards the present case, CopaCogeca enjoys a prominent role since it is an umbrella organisation representing ten million farmers around Europe.[15] More generally, the agricultural sector enjoys a position of prominence among the groups interested in the policy process, since this industry, and in particular the one regarding animal farming attracts a considerable amount of financial support both in Europe and the United States, despite the widely advocated need of a dietary change (Vallone and Lambin, 2023).

2.4 Recognition and relevance

As explained in the above sections, the thesis starts from the assumption that politicians can be characterised as policy-seeking, career-seeking and vote-seeking. On the basis of the understanding of the functioning of the institution, and given the purpose of this research, it is possible to treat career and policy objectives jointly (*see* e.g. Greene and Cross, 2017): this is the reason why the thesis proposed investigates only two out of the canonical three dimensions.

The relevance of the chosen framework (i.e. the informational school of political economy) becomes apparent when considering the possible distortions from third parties more than solely the behaviour of the MEPs. It is arguable that the case under study seems to conform to the prototypical assumption of the theory (You, 2014, pp. 45-50). If the proposal was successful it would have certainly fit the

---

[13] Recently, the access system has changed and the Transparency Portal does not allow for registration anymore. Registration is still required but it should be done through the EP Portal.

[14] https://civio.es/quien-manda/2021/05/20/half-of-all-MEPs-do-not-disclose-any-meeting-with-lobbies/

15 https://copa-cogeca.eu

category of the "disguised transfer" to interested third parties. Moreover, it would have resulted in a labyrinthine set of regulation, typical of the legal regimes where a considerable presence of vested interests are at stake (see Pasour, 1990, with reference to the sugar problem in the Introduction).

Following in the steps of Ibenskas and Bunea, the assumption that *interest group recognition* constitutes a valuable concept for studying the relationships between partisans and interest groups is made. The authors define *recognition* as the attention an organisation receives from a decision maker. While the *relevance* of an organisation may vary depending on the legislation to be passed or the pre-eminence of the organisation on certain issues (the *context*), *recognition* is not necessarily context specific. It "captures a decision maker's interest in an organisation that is motivated by broader and longer-term considerations that are not informed by specific, time-delimited circumstances, and accounts for a more permanent form of organisational relevance and decision-maker attention" (Ibenskas and Bunea, 2020, p. 563).

In line with the cited work of Ibenskas and Bunea (2020), it is argued that an appropriate source of information presenting the desired characteristics are Twitter X followers records. In the presence of a variety of social networks, the decision of using Twitter stemmed from the consideration that it represents an *informational network* (*see* Myers, 2015): a digital forum where it is convenient to share political positions and start campaigns. In addition, MEPs are assumed to be using their time carefully and to inform themselves using organisations that intersect their ideology and provide quality information on key issues (You**,** 2014, Ch. 3). As will be highlighted in the discussion of the results of the Latent Dirichlet Allocation model, this understanding formed the basis for a fortuitous discovery.

Moreover, using Twitter is in line with the current research trends. It has already been shown that MEPs are active on this social and that their posting patterns follow the the agenda of their group of belonging and of the Parliament more in general. The current findings also suggest that MEPs have more incentives in engaging with the social in moments of high visibility (e.g, plenary sessions), when they come from Member States that adopt preferential voting systems, since they can cultivate a "personal brand" and strengthen their relationship with the voters (Daniel *et al.*, 2019, p. 774ff).

3 Bayesian Methods
3.1 Introduction to Bayesian Methods

Since the theoretical background of the more "involved' models presented in this work is Bayesian statistics, it has been judged convenient to present the main ideas that will be applied in the derivation that will follow. Bayesian statistics is a branch of the statistical field based on the work of the Presbyterian Minister Thomas Bayes. The Theorem or "Rule" that brings his name was published posthumously in 1763 in the Philosophical Transactions of the Royal Society. Bayes' Theorem has an intuitive interpretation that may be expressed by the degree to which a subjective belief should change in order to account to a given evidence. In more formulaic terms, it describes how to find the probability $P(A|B)$, that is generally not known, using instead a known probability distribution $P(B|A)$ and some prior assumption about the phenomenon of interest $P(A)$. More formally, the theorem is often stated in the form:

$$P(A|B) = \frac{P(B|A)}{P(B)}$$

Using the rules of probability, the numerator can be expressed as:

$$P(B|A) = P(B|A)P(A)$$

It follows that an alternative formulation for the theorem is:

$$P\big(A|B\big) \propto P\big(B|A\big)P(A)$$

The symbol $\propto$ indicates proportionality, in this way it is possible not to consider the denominator that acts as a normalising constant. Indeed, the denominator can be thought of as the probability of the data (B is what A is conditioned upon, it represents what is known), or equivalently as the likelihood of the data averaged on the prior distribution. Translating the formulas into equivalent statistical concepts, it is possible to affirm that $P(A|B) \propto likelihood \times prior$.

More specifically, the Theorem allows to combine a known probability distribution with some prior assumptions about the phenomenon of interest. From a theoretical perspective, the assignment of a *a priori* distribution has been the most controversial point of debate, and the reason why Bayesian theories have often be

neglected. Still, the formalisation of probability theory, as well as the computational advances, have made the use of Bayesian methods widespread (Robert, 2007).

In order to better understand how the machinery works, the following classical example of Laplace (1763) is proposed: a billiard ball W is rolled on a line of unitary length, the probability for the ball of stopping in any given point is uniform. A second ball O is rolled n times under the same assumptions. A random variable X registers the number of times that O stops on the left of W. If W stopped at p, and it is possible to know only X, what inference can be made on p? (Robert, 2007, p. 10ff). Given the state of things it is possible to assume that p follows a uniform distribution $p \sim U(0, 1)$, and that X is binomially distributed $X \sim B(n, p)$. What follows is only an application of the theorem using the functional versions of the specified probabilities:

$$P\left(X = x \mid p\right) = \binom{n}{x} p^x (1-p)^{n-x}$$

Since it was assumed that p has a uniform distribution, the joint probability of p and X can be expressed as follows:

$$P(x, p) = \int_b^a \binom{n}{x} p^x (1-p)^{n-x} dp$$

Applying the Bayes' Theorem, it is possible to obtain:

$$P\left(p \mid X = x\right) = \frac{\int_b^a \binom{n}{x} p^x (1-p)^{n-x} dp}{\int_0^1 \binom{n}{x} p^x (1-p)^{n-x} dp} = \frac{\int_b^a \binom{n}{x} p^x (1-p)^{n-x} dp}{B(x+1, n-x+1)}$$

The last passage affirms that the resulting probability distribution, called *posterior,* has a closed form, and, precisely assumes the form of a Beta distribution (i.e. $p \sim Be(x+1, n-x+1)$).

Put it differently, it is possible to describe Bayesian statistics as the reallocation of credibility among different probabilities, once some evidence has been observed. Of course, the choice of a restrictive *prior* is liable of distorting the results, and therefore an accurate choice should be made. Nonetheless, it is always possible to opt for flat priors that allows the model to explore a wide range of values (or, in different terms, to follow

the shape of the likelihood of the observations). The Bayesian model that will be presented make use of a number of prior distribution depending on the data (counts, continuous, probabilities etc…). Moreover, the presented model have already been fully developed by others, and the present work simply constitutes an application to a different subject. This is without prejudice to the fact that the choice for the parameters values should and will be justified on a case-by-case basis.

3.2 Conjugacy

As explained in the above section, one of the criticisms moved against Bayesian methods is that the criteria regarding the choice of the prior distribution lack in objectivity. Moreover, the choice of a certain distribution might distort the results, this is the corollary of the updating process described above: indeed, the posterior distribution can be thought of as the least informative distribution with the minimal divergence from the prior that remains consistent with the given data and the constrains imposed on the model.

Three strategies are possible: resorting to a non-informative prior (i.e. a scalar instead of a probability distribution), using distributions having a great entropy, and using conjugate priors. Conjugacy can be defined as follows:

Definition 1 A family of probability distributions $\mathcal{F}$ on a parametric space Θ is said to be *conjugate* if, for every probability density $\pi \in \mathcal{F}$, the posterior distribution $\pi(\theta \mid x) \in \mathcal{F}$.

Put it differently, a prior distribution is said to be conjugate when the posterior distribution is equal to the starting distribution after having developed all the calculations. This ensures that the resulting posterior has a closed form.

3.3 Numerical integration

It is not always possible to reduce posterior probabilities to known probability distributions. In some other cases, the resulting posteriors cannot be sampled. These problems of tractability may be solved by recourse to algorithms of numeric integration. In the present work, different strategies will be leveraged for different models, and they will be presented on a case-by-case basis. In order to better understand the following

parts, in the present section the Metropolis Hastings algorithm (a Markov Chain Monte Carlo method) is introduced.

The class of the algorithms belonging to MCMC aim at creating a Markov chain (i.e. a chain that presents the property that each component only depends on the preceding one), whose stationary (or "long term") distribution is the distribution of interest. More formally, the algorithm can be described as follows:

1) Initialise an arbitrary value θ^{t} where t = 0

2) Update from $\theta^{(m)}$ to $\theta^{(m+1)}$ by:

A) sampling $\xi \sim q\left(\xi \middle| \theta^{m}\right)$

B) choosing $\rho = min\left(\frac{\pi(\xi)q(\theta^{(m)} \mid \xi)}{\pi(\theta^{(m)})q(\xi \mid \theta^{(m)})}, 1\right)$

3) Take $\theta^{(m+1)} = \xi$ with probability ρ if $\rho < u \sim U(0,1)$, otherwise $\theta^{(m)}$.

4) Repeat.

Translating the instructions into words, it is possible to express the algorithm as follows: a random sample (the *current* value) is drawn from the probability density. Then, another distribution called the *proposal* is initialised and a value drawn from it (ξ). The succeeding value in the chain is determined by calculating the ratio proposal/current, and taking the minimum between this ratio and one. The chain is updated by taking the value of the ratio if and only if it is greater than *u,* a number sampled from a uniform distribution between zero and one, however nothing prevents other mechanisms to be used.

In order to make the concept more understandable, it is possible to explain it with a metaphor. Assume that an explorer is on a mountain in a foggy day and he would like to map the of the mountainous region. He only has an altimeter, a notebook and a coin with him. One possible strategy would be that of randomly measuring the altitude at a certain point and noting the value. Then, the measure is taken at a spot some steps farther. In order to decide whether to move or stay in his place, the traveller tosses a coin, and if heads comes up he moves where the second measure has been taken noting the new altitude. In case of tails, he does not move and repeats the process. Such a strategy might not be the most efficient, since the

traveller will either go up and descend in the valleys. Moreover, nothing prevents that he remains stuck in one place for an indefinite time. Still, given the possibility to perform an adequate number of tosses, he should have been able to have a comprehensive view of the landscape: both the peaks and the valleys. Put it differently, the mountainous region is the probability distribution of interest (not known); the current altitude is the starting value of the chain; the second measurement is the sample from a proposal distribution; and the rest of the process is the acceptance and update mechanism.

It should be noted that the proposal distribution might be any possible distribution. However, the choice might affect the time for the convergence of the algorithm. In general, the choice of the Gaussian distribution is justified on the basis of a simplification of the calculations, and also on entropy considerations (*see* McElreath, 2020). In particular, any symmetric distribution (such as the Normal) simplifies the formulae to the form presented above. In the present thesis, only this simplified form will be used whereas the implementation of more involved algorithms will be left to the advanced Python libraries, and will be introduced contextually to the models.

It should be noted that the ratio that constitutes the mechanism of acceptance is an application of Bayes' Theorem. Two conclusions follow: 1) that both numerator and denominator are $\propto likelihood\,x\,prior$, and 2) that if the prior is chosen to be flat (e.g. a scalar such as one), then the quantity that has been called ρ, is a likelihood ratio. Formally, this should still be interpreted as a ratio of *posterior* probabilities, however it both simplifies the calculations and constitutes a strategy when no probability distribution can be assumed *a priori*.

3.4 The Dirichlet distribution

The Dirichlet distribution is a multidimensional probability distribution that may be used to model the probabilities of a series of *n*-outcomes (e.g. the probability distribution of an unfair dice, where some faces have a slightly higher probability of occurring). It follows that the draws from a Dirichlet distribution are vectors of probability summing up to one.

A data structure that affords an intuitive visualisation is the "simplex." Along the edges of the simplex, the probabilities sum up to one. However, there might present areas inside that show a higher probability. Any given point in and along the simplex is identified by a vector of coordinates. The distribution of these values (the

coordinates) depends on the parameter that govern the distribution (what will be called “hyper-parameters” in the paragraph below). There are as many parameters as there are values in the vector.

The following example demonstrates how the high-probability density areas in the simplex change depending on the change of parameters.

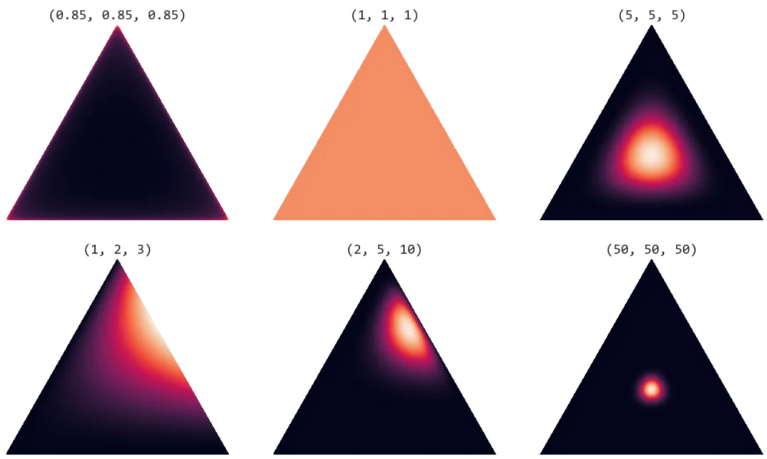


Fig. 1 (Liu, 2019), the change of the probability in a simplex depending on the values of the *alpha* parameters (see functional form below). Consult the reference for a full introductory disquisition on this distribution.

This distribution will be used both in the Latent Dirichlet Allocation model, and in the concluding model to generate vectors of probabilities. Its analytical form is the following:

$$Dirichlet(\theta \mid \alpha) = \frac{1}{B(\alpha)} \prod_{i}^{K} \theta_i^{\alpha_i - 1}, \text{where } B(\alpha) = \frac{\prod_i^K \Gamma(\alpha_i)}{\Gamma(\sum_i^K \alpha_i)}, \alpha = (\alpha_i \dots \alpha_k)$$

This distribution represents a generalisation of the Beta distribution. If a random variable and has mean:

$$E(\theta) = \frac{\alpha_i}{\alpha_0}, \alpha_0 = \sum \alpha_k$$

4 Latent Dirichlet Allocation
4.1 Assessing Copa Cogeca prominence

In the preceding sections the concepts of *recognition* and *prominence* were introduced (*see* 2.4). The first refers to the long-term attention tributed by politicians to certain interest groups deemed to be relevant for the policy process, as such *recognition* is context-independent; conversely, *prominence* depends on the topics being debated, therefore on a context. It is necessary to assess whether Copa Cogeca could be regarded as a "prominent" organisation for the policy at hand.

The choice of Copa Cogeca depends on the fact that, as has been argued in the preceding chapters (2.3), it generally assumes coordinating and leading roles, and it has a long-standing history. Of course, the model has been applied so as to maximise the chances of generating valuable results as will be clarified in the application section (4.3). Moreover, among the other organisations that are active in the same sector, Copa Cogeca routinely publishes policy papers on its website, making it a valuable source of data.

In order to assess the *prominence* of an organisation it is necessary to study whether it has been or was interested in a certain theme, and to what extent. The passage is delicate since it is possible to build upon this basis a rather qualified justification of the use of a variable that links the MEPs with this actor as a predictor of their voting behaviour. That Copa Cogeca is a powerful and respected lobby has already been stated, deriving from this fact that it should be interested in the present case is plausible, but hazardous.

Latent Dirichlet Allocation is an unsupervised hierarchical machine learning model used to find the topics within a text corpus. The model was proposed as an alternative to the more classic term-frequency-inverse-document-frequency (tf-idf) solution (Blei, *et al.,* 2003, pp.994ff). The latter is a method whereby the counts of occurrences of a term in a document is compared to the logarithm of the inverse document frequency (i.e. the counts of documents containing the term in a corpus). The result of the process is a matrix that signals which are the most important words in a document (those that characterise it). Indeed, if *i* represents a word *w* and *j* a document *d,* tf-idf identifies notable words by assigning to them a very low number:

$$w_{ij} = tf_{ij} \log\left(\frac{N}{df_i}\right)$$

Where *tf* is the term frequency for word *i* in document *j; N* is the total number of documents; *df* is the frequency of documents containing the word *i*.

Nonetheless, this method do not reveal anything about the inter- and intra-document statistical structure (Ibid. p.994).

On the other hand, LDA assumes a precise structure for the documents, the model may be expressed by way of a Directed Acyclic Graph (or DAG).

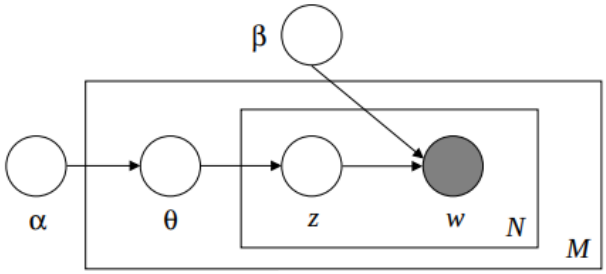


Fig. 2 (from Blei *et al*., *2003)*

The assumption of this model is that each document is a collection of topics (generated by a topic-document distribution), on the other hand each topic is a collection of words (generated by a word-topic distribution). The topic-document distribution is called θ, and it is governed by a hyper-parameter called α. The topic-word distribution will be called ϕ, and the hyper-parameter governing it will be called β (this is a variation from the DAG, however it will make the ensuing derivations easier to follow). In the above DAG, z represents a topic; w represents a word; N represents the set of all words in a given document; M represents a given document i a corpus; the boxes represent "for-cycles" (i.e. programming structures that iterate over a collection of items).

In more discursive terms, it is possible to affirm that the model assumes that something is known (the words that are observed) and that from this known it should be possible to make an inference about the unknown (the topics across a set of documents, a corpus). It should be noted that the unknown, the topics, are characterised as set of words whose probability of being together is comparatively higher than other sets of words.

The topic-document and topic-word probability distributions are generally chosen to be Dirichlet distributions, and since the process is iterative it will be assumed that the topics and words are sampled from time to time from discrete Multinomial distributions. Therefore, the model can be written synthetically as follows:

$$
\begin{aligned}
w_i \mid z_i, \phi^{(z_i)} &\sim \mathcal{M}(1, \phi^{(z_i)}) \\
\phi &\sim Dirichelet(\beta) \\
z_i \mid \theta^{(d_i)} &\sim \mathcal{M}(1, \theta^{(d_i)}) \\
\theta &\sim Dirichelet(\alpha)
\end{aligned}
$$

Expressing it in words, it is possible to interpret the model as follows: the probability of observing a word in document *i,* is sampled from a Multinomial distribution parametrised by ϕ, the topic-word Dirichlet distribution depending on the hyper-parameter β. For each word in a document, the topic and the topic-words distributions are given (this is reflected in the structure of the DAG, where the edges connecting the nodes represent conditional probabilities). Still, for a topic to be given, it should be sampled from its distribution, and, specifically, for each document *i,* from a Multinomial distribution parametrised by the document-topic Dirichlet distribution θ depending on the hyper-parameter α.

4.2 Deriving the posterior probability

In order to be able to sample from this hierarchical model, a classical method is Gibbs sampling, a special case of the Metropolis Hastings algorithm, where conditional distributions are used instead of proposal distributions. An example that illustrates the process is the following, assume joint probability density of two random variables X and Y is known:

$$f(x,y) \propto \binom{n}{k} y^{x+\alpha-1}(1-y)^{n-x+\beta-1}$$

$$\text{with } x = 0,1,....n \text{ and } 0 \leq y \leq 1$$

If the distribution of interest was $f(x)$, it could be possible to consider the two conditional distributions that is possible to derive directly from the joint by suppression of independent terms. These distributions are:

$$f(x \mid y) \sim Bin(n,y)$$

$$f(y \mid x) \sim Beta(x+\alpha, n-x+\beta)$$

As can be seen, to generate samples for X it is possible to sample from the conditional distribution of y (since *n* is given), and it is possible to use this value to sample from the conditional distribution of X (to see an application of this derivation and other examples refer to Casella and George, 1992, pp. 168ff).

The same reasoning might be applied to the sampling of the distribution involved in the Latent Dirichlet Allocation model. To be sure, the problem at hand is under all respects similar to the example proposed. In the case of the LDA, the words are assumed to be observed and known, instead the topics of which a text is

composed are not. Therefore, the interest lies primarily in determining the conditional probability of the topics, or $P(z \mid w)$.

While it is possible to find several complete derivations of the full model elsewhere (see Koks, 2019), in this work another path will be proposed. Indeed, there are at least two ways that might be followed: expansion and cancellation of terms (involving lengthy calculations as in Koks), or making a probabilistic argument (Griffiths and Steyvers, 2004, p. 5229, that will be also the main guide for the ensuing paragraphs).

The rationale behind this alternative methodology is that it is more elegant, intuitive and efficient (Ibid.) As will be clarified below, the reason for this increased efficiency is mainly due to the fact that in order to derive the topic probability, it will be assumed that one of the distributions of the model had been suppressed, and that the formula that will be arrived at will only deal with counts. This suppression is without prejudice as regards the objective of the derivation itself (i.e. finding $P(z \mid w)$).

Assume that the model above does not feature at all the topic-word distribution ϕ, and the document-topic distribution θ. In this scenario, the probability distributions left would be only the topic and word distributions, both Multinomial. The joint probability distribution of the words and topics can be expressed by common probability relationships as:

$$P(w, z) = P(w \mid z)P(z)$$

It is possible to enrich this general step with a deeper knowledge about the state of affairs. The above equation conditions the probability of a word to a topic, however it is not true that at any moment the topics are all unknown. As can be recalled from the introduction to the Metropolis Hastings algorithm (of which Gibbs sampling is a special case), there is an initialisation step for these MCMC methods to produce chains converging to the distributions of interest. Therefore, even at the first iteration, it is possible to know what are the topics known (even though at such an early stage they might be not representative of the real topics). This knowledge can be expressed, momentarily by a vector ζ.

$$P(w, z) = P(w \mid \zeta)P(z \mid \zeta)$$

In order to give meaning to ζ, it is necessary to develop further the reasoning. It was assumed that the topic-word and document-topic distributions were not existent. Mathematically, this would be equal to saying that those two distributions had been "integrated out." Note from the distributions listed above, that the probability distribution of *w* only depends on θ, and that the topic distribution only depends on ϕ. Therefore, assuming that the two Dirichlet distributions were not existent is mathematically equivalent to assuming that the cause of non-existence was integration:

$$\int_{\phi_k}\int_{\theta_d} P(w,\phi_k|\zeta)P(z,\theta_d|\zeta)d\phi_k d\theta_d$$

In the writing above, *k* represents a given topic in a list of topics ranging from{1….k}, and *d* represents a document in a list of documents ranging {1….d}. Now it is possible to give full meaning to ζ because the distributions that had been suppressed appear again in the equation: it is possible to state what is known exactly at each time (because defined in advance or given as input data). Therefore, the above expression is equivalent to:

$$\int_{\phi_k}\int_{\theta_d} P(w_{id},\phi_k|\alpha,\beta,z,z_{-id})P(z_{id},\theta_d|\alpha,\beta,z_{-id})d\phi_k d\theta_d$$

While the above notation is more complete, it makes the derivations more burdensome to follow and will be made lighter in the next steps. However, it expresses the state of knowledge at a given point in time: for each word *i* in a document *d*, the hyper-parameters and all the topics, both the current (i.e. the one associated with the word being examined) and the the remaining ones, are known. On the other hand, when examining the probability of a topic for word *i* in document *d*, both the hyper-parameters and all the other topics but the current one are known. For the sake of brevity, the ensuing derivations will revert to ζ.

A remark to be made on the words is the following: not every single word of the original documents forming a corpus must be taken into consideration. For a smoother convergence of the model it is necessary to clean the textual data. This can be achieved, after having eliminated the stop words, by creating a vocabulary (containing each word one time), and a count matrix (containing for each row representing a document, how many times each word in the vocabulary, the columns, repeats itself). Since the model is applied on such a matrix, its output (the

topics) can be thought of as the words who tend to be associated with the highest probability within a collection of texts. This probability distribution is estimated on the counts of the words, and it is a discrete probability distribution (See in particular Koks, 2019, ch. 2-3).

To show why the latter proposition holds true, a suitable argument might be based on conjugacy (3.2). In order to make the derivation simpler, it is possible to focus on one integral at a time. Starting with the one in ϕ, two things are apparent: 1) that there is a joint probability, and 2) that if this probability is written, equivalently, in conditional form, then a Dirichlet distribution (the topic-word) gets multiplied by a Multinomial distribution (the word distribution). However, the Dirichlet distribution is conjugate prior of the Multinomial. It follows that the resulting distribution is again a Dirichlet distribution, updated on the basis of the observed words. Developing the calculations, sheds some light on the exact nature of the updating process:

$$P(w \mid \zeta) = \int_{\phi_k} P(w, \phi_k \mid z) d\phi_k = \int_{\phi_k} P(w \mid z, \phi_k) P(\phi_k) d\phi_k$$

The above formula, applies to all the words belonging to {1....w}. Put it differently, it computes the marginal probability of the data fixed a topic (i.e. all the possible word distributions given a topic). In order to compute the probability of a specific word given a specific topic, it is necessary to rewrite the integral so that the above quantity gets multiplied by the distribution of word *w* given topic *k:*

$$P(w \mid \zeta) = \int_{\phi_k} \phi_{kw} P(w \mid z, \phi_k) P(\phi_k) d\phi_k$$

Expanding in functional form produces the following results:

$$P(w \mid \zeta, \phi_k) = \int_{\phi_k} \phi_{kw} \frac{1}{B(\beta)} \frac{\Gamma(\sum_i w_i + 1)}{\prod_i \Gamma(w_i - 1)} \prod_i^k \phi_i^{(\beta_i - 1)} \prod_i^k \phi_i^{(w_i)} d\phi_k$$

$$\propto \int_{\phi_k} \phi_{kw} \prod_i^k \phi_i^{(\beta_i + w_i) - 1} d\phi_k$$

$$\propto \int_{\phi_k} \phi_{kw} Dir(\beta_1 + w_1, \beta_2 + w_2 \ldots \beta_i + w_i) d\phi_k$$

The meaning of the latter expression is that to each hyper-parameter (originally a vector of dimension *n* uniformly initialised), a certain quantity, the counts for specific words, is added. The Dirichlet distribution, at this point, represents the update in the belief that a certain topic is being observed based on a sequence of counts corresponding to distinct words. As shown above, this results in a simplex concentrated in different points on the basis of the words, thereby allowing it to be connected to the identification of a topic. Using the relationship between the Beta and the Dirichlet distribution, it is possible to interpret the above equation as the mean of a Beta distribution representing the best guess at a word given a topic:

$$\propto \int_{\phi_k} \phi_{kw} Beta(\beta_{kw} + u_{kw}^{-id}, \sum_{j \neq w} u_{kj}^{-id} + \beta_{kj}) d\phi_w$$

Therefore,

$$P(w \mid \zeta) = \frac{\alpha}{\beta} = \frac{u_{wk}^{-id} + \beta_w}{\sum_{j \neq w} u_{jk}^{-id} + \beta_j}$$

Where *u* is used only to stress that what is added to the hyper-parameter is not a word proper (a "string"), but a count corresponding to a word. It is also emphasised that this count does not contain the word being observed (i.e. word *i* in document *d*).

The same steps may be applied to the second part of the integral. Since the reasoning remains the same, the procedures are presented in a more concise way below:

$$P(z \mid \zeta) = \int_{\theta_d} P(z_{id}, \theta_d \mid \alpha, \beta, z_{-id}) \theta_d$$

As before, the interest lies in:

$$\int_{\theta_d} \theta_{dk} P(z_{id}, \theta_d \mid \alpha, \beta, z_{-id}) \theta_d$$

$$\propto \int_{\theta_d} \theta_{dk} Beta(\alpha_{dk} + o_{dk}^{-id}, \sum_{j \neq k} o_{jd}^{-id} + \alpha_{jd}) d\theta_d$$

$$P(z \mid \zeta) = \frac{\alpha}{\beta} = \frac{o_{kd}^{-id} + \alpha_k}{\sum_{j \neq k} o_{jd}^{-id} + \alpha_j}$$

Having derived the two means of the Beta distributions, now it is possible to answer the original problem. Given a corpus made of texts some words are observed, however an inference about the most probable collection of words unified in a topic should be made. In statistical terms the object of interest is the following, unknown distribution: $P(z \mid w)$. By applying Bayes' Rule, it is possible to write the conditional as a joint, normalised by a quantity in the denominator. However, the resulting expression is the form a posterior distribution, which is proportional to the numerator only (the topic distribution is being integrated out in the denominator, so that $P(z \mid w)$ is not dependent on it).

$$P(z \mid w) = \frac{P(w, z)}{\sum_z P(w, z)} \propto P(w \mid z) P(z)$$

The resulting formula is now composed of known elements only, precisely those that could be found in the preceding steps (indeed, while not written explicitly, $P(z)$ and $P(z \mid \zeta)$ are the same quantity, since for the convention that has been proposed ζ is a dummy vector that represents the known elements at any given time, depending on the distribution that is being studied). Finally, the answer to the problem can be given in functional form:

$$P(w \mid z) P(z) \propto \frac{u_{wk}^{-id} + \beta_w}{\sum_{j \neq w} u_{jk}^{-id} + \beta_j} \frac{o_{kd}^{-id} + \alpha_k}{\sum_{j \neq k} o_{jd}^{-id} + \alpha_j}$$

It should be noted that there is a discrepancy between the last formula and the derivations above: the hyper-parameters have only one index in the final formula. This is because in the application of the model, it has been chosen to initialise α and β uniformly, as a vector of dimension n repeating the same value. In principle, if more was known about the data, it could be possible to initialise a different value for each hyper-parameter of the vector. The result would be a Dirichlet distribution either more sparse or concentrated in some points (as shown

4.3). In the application of the model, the correct initialisation has been found tentatively after several tries.

4.3 Application

The model has been applied to a curated dataset of policy papers and tweets of CopaCogeca. Thirty-five policy papers spanning the years 2018-2023 were manually retrieved. The reason why this part of the process was not automated is that it was necessary to check the quality of the files. Automating would have allowed to retrieve more data, however it has been chosen that in order to make the model converge on the topic of interest, it would have been necessary to restrict the time-frame so that it could be both representative of the period in which the vote took place, and broad enough to allow for a comprehensive view of the evolution of the topic.

It was necessary to retain the temporal information for the correct application of the methods. Exploring the files publicly available on the website of Cop (in the "Publications" section), it is possible to note that the date in which they were published on the website and those that can be read within the files are matching in almost all cases. When discrepancies were found, the earliest of the two dates was chosen since it has been interpreted to mean that the positions expressed had already been developed.

Along with these official data, also the tweets from the same period were considered. The two bases of data were merged and the strings concatenated when a tweet and a policy paper were published on the same date (in both cases it was possible to know day, month and year of publication).

Text processing has been performed following the conventional steps: stop-words, in-text dates and non alphabetical characters were removed. As common practice, the text was converted in lowercase, tokenised (reduced to discrete units), and lemmatised (the words were brought back to their roots, e.g. "going" was turned to "go", "students" to "student" etc…). Finally, with the hep of the available tools, it was possible to remove the words possibly insignificant in the text (either because they appeared with a frequency of above 95% or below 3%).

After these cleaning steps, the text was converted in a matrix where each row corresponded to a policy paper or tweet, each column to a word, and the cells contained the counts of the word per entry. The words, however, were taken from a vocabulary and not the raw text, in this way each unique word was repeated only

one time in the columns. This structure was particularly useful, for it made possible linking the output to the "distribution" in time of the topics found.

It should be remarked that the model being applied is an unsupervised technique of machine learning. The steps described, from the choice of the material to their handling, were taken as a precaution in the hope that the output could be related to the purposes of this work.

Considering the length and general structure of the policy papers, as well as the tweets, an appropriate number of topics was assumed to be thirty (several empirical trials have shown that a higher number of topics resulted in less interpretable results with the same hyper-parameters). The alpha and beta hyper-parameters were initialised at the values of 0.01 and 0.5 respectively, underscoring the idea that the variability of topics within a document should be more than the variability of words defining a topic.

The output of the model consists of a list of ten topics, a document-topic matrix and a topic-word matrix. As explained, the dates were used as document titles so that it was possible to relate the output to the diachronic development of each topic. In this sense, the results of the model are broader than the scope of the paper, since they represent some of the topics that have characterised CopaCogeca activity during the time-frame considered. The table in Appendix I presents the keywords associated with each topic, and an interpretation of their meaning that is given through the attribution of a title to each of them.

As can be seen, the fifth topic, characterised by the words "denomination cecinestpasunsteak plantbased know meat marketing imitation renew meatdenominations cultural", seem to be particularly relevant and most likely deriving from Twitter X data, since the presence of the word "cecinestpasunsteak" seems to be part of an hashtag. The distribution of this topic, reveals that Copa Cogeca was attentive to the issue in the neighbourhood of the vote.

The dotted red line signals the day of the vote of the proposal under study, interestingly in correspondence of the major peak in the distribution. It is also apparent that Copa Cogeca had started dealing with the issue some months before the vote.

Upon a qualitative investigation, aided by the keywords provided by the Latent Dirichlet Allocation model, it can be concluded that the interest of Copa Cogeca was systematic and not generic. In the "campaigns" section of the official

website of the organisation it is possible to find the “Ceci n'est pas un steak!” campaign that started on 15/10/2020 and ended on 23/10/2020.[16] The existence of the campaign has not been found in the papers consulted for the writing of this work, and seem to be a novel discovery. Moreover, an internet search of the manifestos revealed that other organisations took part in the campaign, notably: EFFAB, AVEC Poultry, CLITRAVI, and UECBV. These findings will be incorporated into the logistic regression model that will be proposed in the following section.

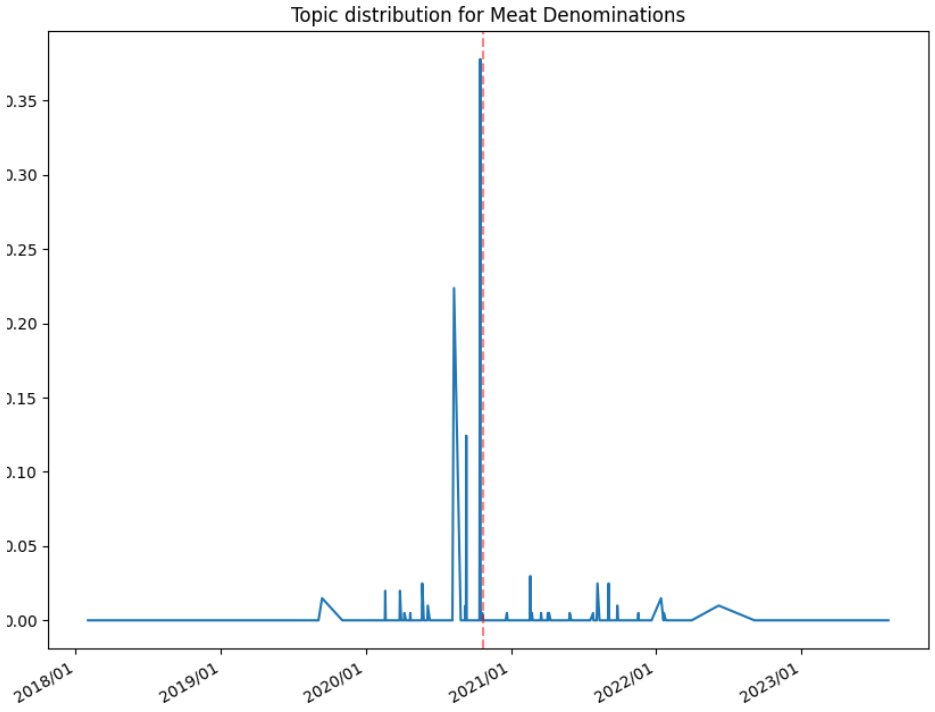


Fig. 3 Distribution of topic 5 with respect to the time of the vote (dotted red line). It seems that the topic was discussed more frequently at a neighbourhood of the vote.

While a qualitative research could have led to similar results, the approach adopted presents several advantages. On the one hand, it was possible to find twenty-nine topics other than the one of interest that seem to be generally interpretable and related to Copa Cogeca’s activities. On the other hand, the model made possible to visualise the topic distribution in time. A qualitative research would have most likely identified the existence of a campaign, without arriving at the level of detail presented in this case. Indeed, it is possible to generalise beyond the starting and ending dates of the campaign, as it would seem that Copa Cogeca had been interested in the discussions surrounding the marketing of meat-substitutes some months before the month of the vote.

[16] https://copa-cogeca.eu/campaigns

5 A Multi-intercept regression model
5.1 Introduction

Latent Dirichelet Allocation has allowed to substantiate the claim that a set of organisations centred on CopaCogeca was particularly interested in the legislative proposal under study. In order to clarify the relationship between these interest-groups and the voting behaviour of the MEPs, it is necessary to develop another model on an alternative base of data.

As explained in the theoretical part, current research has pointed out several aspects of the legislative behaviour of the MEPs that are apt to be translated into a quantitative language. Incorporating the critique that the informational school of economics remains too focused on an individual level, it is argued that MEPs are certainly prone to be influenced by an interest group on the basis of their idiosyncratic considerations, however they are also part of broader organising structures. Therefore, the belonging to a committee or a party should be taken as an information reflecting the natural grouping of the "society" under study. In the present case, the Envronmental (ENVI) and the Agricultural (AGRI) committees are relevant (since, as explained ENVI gave an opinion on the subject and AGRI was the competent committee to examine the proposal).[17]

On the other hand, it would be useful to identify some variables in order to represent the priorities of the MEPs as regards their willingness to innovate the current policies and to be re-elected. With regards to the first aspect, it is argued that such a willingness might be measured by MEPs recognition of the relevant interest groups. As pointed out above, it is reasonable to assume that that the long-term commitment of politicians to a certain political view should correspond to a net of "informers" that are deemed to be relevant by the MEPs when they are carrying out their work: time is a precious resource that MEPs are expected to use in the most efficient manner (*see* e.g. You, 2014). In line with the current studies, the measure for this linkage has been reputed to be the follower count of the Twitter (X) profiles of the relevant lobbies.

However, it is also true that MEPs have to deal with a national dimension. In order to test for the effect of the national constituencies of belonging, a measure of leaning towards more urban or rural matters has been retrieved from the Chapel Hill

[17] After several trials, participation to either one or the other committee has been registered in a single predictor variable. While this has not had any influence on the results, it has made the summary more readable.

Expert Survey Dataset.[18] This index, ranging from zero to ten, highlights the partisans that come from national political groups that privilege either urban (zero) or rural (ten) interests. Career-seeking attitudes, can be spotted by considering both the constituency and the election system of the MEPs. Some countries adopt a system of indirect election, where the national parties choose who to elect for the European Parliament, others a system where the citizens can express directly their preference. It is reasonable to assume that in the second case the MEPs should be more consistent with the orientation of their electoral base (in other words that there is a positive correlation between a binary value assuming value one when the system is open, and the index measuring urban-rural leaning).

Given the considerations made so far, it would seem that an appropriate model to study the voting behaviour of the MEPs should be a logistic regression where the variables described above should be the independent predictors, and the outcome of the vote the dependent variable. Moreover, it is reasonable to assume that each European political group should have a different intercept.

The model proposed will not be fitted through Bayesian methods, since this would require an high number of distributions, long-form calculations and the implementation of an advanced sampling method (the No U-Turn sampler). To be sure, it is possible to simplify both the derivations and the sampling process significantly: PyMc, a Python library for Bayesian statistics, automatically samples from the posterior distribution avoiding coding complexities. Since even the frequentist version of the model requires numerical methods for fitting (automated by Python *statsmodels* library), it seemed that the Bayesian model would have resulted in an unnecessary complication. For a dataset such as the one of the present thesis, this choice seemed not to be producing a sensible difference in the results.

The data are briefly summarised in the following table:

Variable Name	Description	Variable Type
urban_rural	Index of urban-rural leaning of national parties derived from the CHES dataset.	Numerical, ranging continuously from zero to ten.

[18] The CHES dataset is a data source curated by experts founded by several academic institutions who survey the party positioning across Europe and America (South and North). The data is freely accessible from their portal: https://www.chesdata.eu.

Variable Name	Description	Variable Type
politicalGroup	MEP belonging to European parliamentary group.	Categorical (party names). In the implementation of the model it will be treated as binary (see the specification of the model).
isOpenList	Signals whether an MEP comes from a national party	Binary.
isCommitteeMember	Signals whether an MEP is part of the AGRI or ENVI committees.	Binary.
copa_cogeca_net	Signals whether an MEP follows at least one of the Twitter pages of Copa-Cogeca, EFFAB or AVEC Poultry. CLITRAVI does not have a Twitter page. The sectoral pages of Copa Cogeca have been considered, specifically: Copa Cogeca Meat and Copa Cogeca CAP	Binary.
vegetarian_union_follower	Signals whether an MEP is follower of the Vegetarian Union page (main stakeholder from the opposite side of Copa Cogeca).	Binary.
VOTE	Vote cast in the plenary voting session.	Binary (the abstained were not considered).

5.2 Deriving the model

Logistic Regression is a Generalised Linear Model (i.e. a model where the parameters of the likelihood function have been replaced by a linear model). In contrast with the previous model, it represents a supervised classification method.

In order to derive the model, it is necessary to introduce the concept of "odds." While probabilities are numbers ranging between zero and one that represent the proportion of the occurrences of an event over the total number of trials; odds are defined as the proportion of the probability of an event occurring,

divided by the probability of the event not-occurring. As such, odds are number not limited between a given range. The latter property is particularly useful for fitting a linear model, whose output is not constrained in a given interval.

Logistic regression uses the logarithm of the odds, or the "logit" or "sigmoidal" function, characterised by an s-shape. If p is the probability of the vote being favourable, y the dependent variable, and X are the predictor variables:

$$logit(p\,|\,X = x) = log\left(\frac{p_i}{1 - p_i}\right)$$

The above, as anticipated should be equal to a linear model, that in this case should contain a general intercept, and group level intercepts that signal the model whether an MEP belongs to a certain political group. The dummy D takes on value zero when the partisan is not the member of the corresponding group, and value one otherwise:

$$logit(p\,|\,X = x) = log\left(\frac{p_i}{1 - p_i}\right) = \alpha_0 + \alpha_{g1}D_{g1} + \alpha_{g2}D_{g2} + \ldots + \beta_1 x_1 + \ldots + \beta_n x_n$$

Taking the exponent of both sides it is possible to find the probability of the dependent. variable:

$$p_i = \frac{e^z}{1 + e^z}$$

$$\text{Where } z = \alpha_0 + \alpha_{g1}D_{g1} + \alpha_{g2}D_{g2} + \ldots + \beta_1 x_1 + \ldots + \beta_n x_n$$

As recalled, the fitting method is based on maximising the log-likelihood function:

$$\mathscr{L}(y_i) = \prod_{i=i}^{N} p_i^{y_i}(1 - p_i)^{1-y_i}$$

$$\log\left(\mathscr{L}(y_i)\right) = \sum_{i=1}^{N}\left[y_i \log p_i + (1 - y_i)\log(1 - p_i)\right]$$

Since it has been shown that:

$$p_i = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-(\alpha_0+\alpha_{g1}D_{g1}+\alpha_{g2}D_{g2}+\ldots+\beta_1 x_1+\ldots+\beta_n x_)}},$$

the Maximum Likelihood strategy to fit the model (i.e. to find the most appropriate values for the intercept and the coefficients), is maximising the log-likelihood function. The algorithm used for this model is the Newton-Raphson, that involves the following steps.

1) An initial guess x_0 for a root of the objective function $f(x)$ is made.

2) The tangent line of the function at the specified values is computed: $y = f(x_0) + f'(x_0)(x - x_0)$

3) The next value in the series is found by studying where the intercept intersects the x-axis. Developing the calculations by setting the tangent line equals to zero yields: $x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$

4) The process is repeated iteratively with the new values taken as guesses. The s*tatsmodels* library provides for an automated halting system that calculates when the differences between successive approximations descends under an internally defined threshold for all the parameters (the system also checks on the log-likelihood change rate, interpreting negligible change as convergence).

As can be inferred from the functioning of the model there might be convergence problem when the derivative of the function does not exist. Moreover, convergence time depend on how far is the initial guess from an actual root of the function. In the present application, these problems have not surfaced.

Another result of interest is the formula for calculating the relative effects of the predictors on the odds of the event occurring. In order to simplify the calculations, it will be assumed that the model has only one regressor and one intercept. As before, the relationship that links the odds and the linear model is $\frac{p}{1-p} = e^{\alpha+\beta x}$. Now it is possible to define the proportional odds as the quantity that describes the change of the odds when the regressor is augmented by one unit. Developing the calculations, it is possible to arrive at the conclusion that relative effects are the exponential of the coefficient of the regressor of interest:

$$\frac{e^{\alpha+\beta(x+1)}}{e^{\alpha+\beta x}} = \frac{e^{a}e^{\beta x}e^{\beta}}{e^{\alpha}e^{\beta x}} = e^{\beta}$$

5.3 Application

The logistic model described above has been fitted on the vote data, split in a training dataset comprising 80% of the observations, and it has been evaluated on a test dataset comprising the remaining 20% of the observations. The fitting has produced the following results:

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-1.9216	0.370	-5.193	0.000	-2.647	-1.196
C(politicalGroup, Treatment(reference=baseline_group))[T.ECR]	1.4376	0.418	3.435	0.001	0.617	2.258
C(politicalGroup, Treatment(reference=baseline_group))[T.EPP]	0.7060	0.287	2.458	0.014	0.143	1.269
C(politicalGroup, Treatment(reference=baseline_group))[T.GUE/NGL]	0.2775	0.505	0.549	0.583	-0.713	1.268
C(politicalGroup, Treatment(reference=baseline_group))[T.Greens-EFA]	-2.6329	1.037	-2.540	0.011	-4.665	-0.601
C(politicalGroup, Treatment(reference=baseline_group))[T.ID]	3.6621	0.775	4.725	0.000	2.143	5.181
C(politicalGroup, Treatment(reference=baseline_group))[T.Non attached]	-0.8582	0.797	-1.077	0.281	-2.420	0.704
C(politicalGroup, Treatment(reference=baseline_group))[T.Renew]	0.6598	0.350	1.888	0.059	-0.025	1.345
urban_rural	0.2066	0.063	3.262	0.001	0.082	0.331
isOpenList	0.1988	0.225	0.882	0.378	-0.243	0.640
isCommitteeMember	-0.1526	0.278	-0.550	0.582	-0.697	0.391
copa_cogeca_net	-0.9966	1.198	-0.832	0.405	-3.344	1.351
vegetarian_union_follower	-42.4470	5.8e+09	-7.32e-09	1.000	-1.14e+10	1.14e+10

It should be noted that a “treatment” was applied to the data, i.e. it was possible to specify a baseline group relative to which the varying effects of different groups were calculated. In the present case, the baseline was chosen to be S&D because the proposal came from MEP Eric Andrieu who was affiliated with that party. The choice of a baseline might be useful for interpretative purposes, and does not affect the predictive capabilities of the model.

With regards to statistical significance, it may be noted that the p-value predictors are all above the 5% threshold, except for the *urban_rural* index from the CHES dataset. The general intercept (incorporating the effect of the S&D party) is statistically significant, along with the intercepts of ECR, EPP, Greens/EFA and ID. All of the parties have positive intercepts except for the Greens and Renew, indicating a tendency to vote favourably when compared to the baseline party. Put it differently, it would seem that the proposing party displays a statistically significant tendency towards an unfavourable vote, while other parties, and most notably ID, have an greater propensity towards favourability when compared to S&D.

Building on the results presented in the above section, the following tables compares the relative effects between groups. As could be inferred from the coefficients, ID seems to display the highest relative effect when compared to the

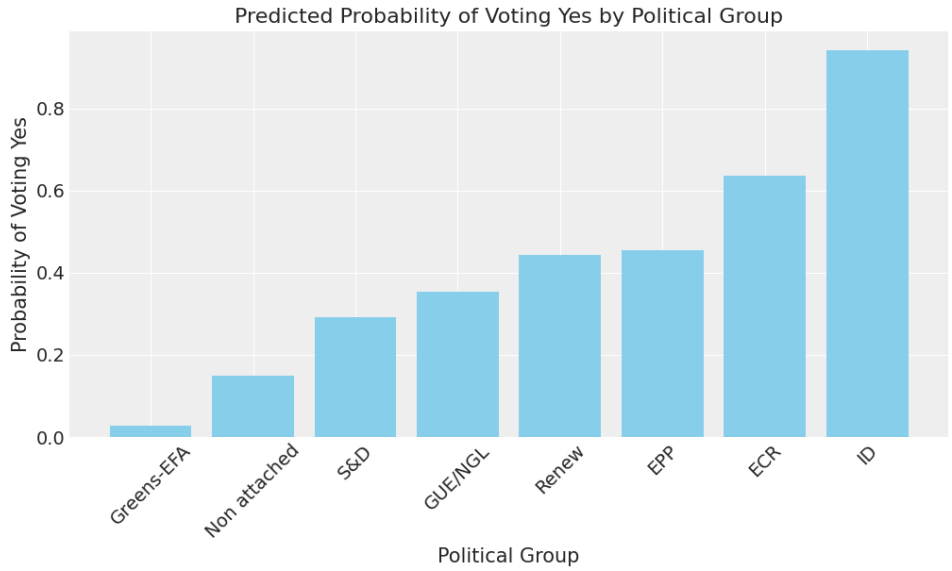


Fig. 4 Probability of positive vote depending on political group.

baseline. The interpretation is that a unit increase in the coefficient considered, produces an increase in the log-odds of observing a favourable vote that of a magnitude equal to the value of the exponentiated-coefficient.

Group	Coefficient	Relative effect
S&D	-1.9216	0.14
ECR	1.4376	4.21
EPP	0.7060	2.03
GUE/NGL	0.2775	1.32
GREENS	-2.6329	0.07
ID	3.6621	38.94
NON ATTACHED	-0.8582	0.42
RENEW	0.6598	1.93

In order to display how the probabilities of a positive vote change depending on the group, it is also possible to run the fitted model on a new dataset made up of only the unique political groupings, associated with the mean values of the predictor variables used in the general model. In this case, the logistic regression that was fitted in the preceding step, is given the mean value of the covariates as an input, and returns, per each group, the probability of observing a positive vote.

With a pseudo R-squared of 0.25, the model displays an acceptable fitting. Logistic regression models are usually evaluated on the basis of certain metrics that are listed and described in the following table:

Metric	Value
Accuracy	0.71
Precision	0.71
Recall	0.55
F1	0.62

The Area Under the Curve (AUC) score is of 0.77. This measure indicates the area under the Receiver Operating Curve, that plots the relationship between true positives (y-axis) and false positives (x-axis). The more the curve leans towards the upper-left corner of the plane, the less are the false positives when compared to the true positives, and the better the model.

While the results from the model are insightful, it is not straightforward to give to them a casual interpretation for two reasons: (1) most of the predictors are not statistically significant, and (2) the model is not apt for providing such interpretation.

As regards (1), it should be noted that statistical insignificance does not necessarily imply that the single predictors are not valuable, instead it should be interpreted as meaning that they are not reliable sources of information for the given vote, when the model is specified in the way that has been presented. There might be alternative models that incorporate the same variables, or models presenting the same variables excluding some of them, that predict the voting outcome more efficiently.

However, the reason why the search of an alternative model is not pursued in the present thesis is that even if a better model could be found, it would not have been sufficient. Indeed, the research question includes an element that requires to make a causal inference on the vote. Given a vote, it is necessary to determine whether the *cause* was either a policy or career-seeking behaviour of the MEPs. With reference to (2), it should be noted that the regression made no assumptions on the causal structure of the phenomenon, limiting itself to studying the correlations between variables and vote. Therefore, the positive coefficient for *urban_rural* should not be interpreted as a causal inference of the constituency on the behaviour of the partisans. The relative effects (presented in table) might give a more nuanced and yet not conclusive understanding: indeed, it is widely acknowledged that a

regression model might predict perfectly and provide poor causal understanding, while a causal model might provide poor predictions and provide valuable causal inferences (McElreath, 2020, Ch. 6).

6 From association to causation

6.1 Introduction

The challenges regarding the choice of the type of inference required to answer the research question of the present work might be ascribed to the class of "inverse problems." While it is generally straightforward, knowing the cause, to predict the effect, the present study is centred on deriving the plausible causes of an observed behaviour, i.e. a certain vote held in the European Parliament. There exist two categories of inverse problems: *reconstruction* and *parameter estimation*. The first consists of determining the input that produced a certain output; the second, followed in the present work, concerns the estimation of the parameters of a model (Waqar *et al*., 2023, p. 3ff).

Bayesian statistics is particularly apt for the study of *applied* inversion problems. Theoretical inversions might not necessarily account for the noise in the data. Indeed, when dealing with real-world cases the existence of a solution, and its uniqueness (that might generally be proven under theoretical scenarios) are not granted (Ibid.) As illustrated in the above sections, Bayesian methods allow for an incorporation of the prior knowledge of the researcher and provide a posterior distribution that incorporates the uncertainty in the resulting estimates.

Logistic regression, and classical statistical models based on the association between predictors and outcome, do not represent a sufficient means of analysis for the present case. It is possible to make sense of this inadequacy by considering the following anecdotal example: suppose that an hospital would like to predict whether the rate of incoming patients for the next day will be high or low using a register of arrivals and a vector of predictor variables. A logistic regression model would provide, depending on the regressors, reliable indications as to what should be expected. However, the model does not contributed *per se* the knowledge about what exactly produces a considerable number of arrivals (i.e. it does not include a representation of how the change in one independent variable affects the dependent variable), since it does not include explicitly a causal structure. It would certainly be possible, as has been done in the present case, to study the relative effects of the predictors, and yet this approach remains insufficient since it lacks a broader understanding of the generative process underlying the observations.

As noted by Pearl (1995, 2012, and in particular 2010 pp. 1-10), causal assumptions identify invariant relationships despite the change of external conditions. This peculiar typology of questions cannot be addressed neither through a joint probability distribution, nor solely through the concept of statistical dependence dependence. Indeed, while the first is clearly unsuited to describe a statement such as "the symptoms do not cause sickness" (positive correlation), the second limits the beholder understanding to the fact that when there is sickness, symptoms are also commonly found $P(\text{sickness} \mid \text{symptoms})$. It follows that the main difficulty in this typology of inference is that causal assumptions are not verifiable, unless it is possible to put them to the test in an experimental setting. Even Bayesian statistics, in which is possible to encode prior beliefs explicitly, does not necessarily provide *per se* a framework to study these problems: the sensitivity of the model to the priors tend to diminish as the sample size increases, whereas the sensitivity to causal assumptions remains invariant (Ibid. p. 3).

It is possible to represent the observations (the vote) as the result of an unobserved (and likely unobservable) process. As assumed by the literature, partisans are agitated by two main concerns regarding their career progression: on the one hand they are policy seekers, they would like to innovate the current policies to gain power and reputation within their institution; on the other hand, they should prioritise their constituencies in order to be re-elected. These examples of rational behaviour will be called "strategies" Therefore, on the basis of what has been said until this point, it should be correct to state that the observed votes have been produced by certain strategies, having different *probabilities*. The research question might be reframed as the computation of the probabilities that underly those strategies (i.e. asking with what probability politicians were policy seekers instead of vote seekers).

In order for the model to be applied, it is necessary to define *a priori* some plausible strategies. While some of the determinants of MEPs behaviour have been found to be statistically insignificant, this is without prejudice to the use of the same variables in the new model. As remarked in the above paragraph, the purpose and hypotheses of the models are different. Certainly, the strategies that will be proposed might not be the ones that have "caused" the observations. Nonetheless, it is possible to justify the approach adopted on the basis of the following considerations: (1) the strategies defined are based on what has been found significative in the literature as regards MEPs' behaviour; (2) in principle other possible strategies might be included in the model and their relationship

studied; (3) the results from the logistic regression are incorporated into the hypotheses of the new model. It is argued that the three justification proposed contribute to a valid model. As will be discussed in the conclusions, it is in the interplay of the three models proposed that the research question finds a complete answer.

6.2 A *do-calculus* model introduction

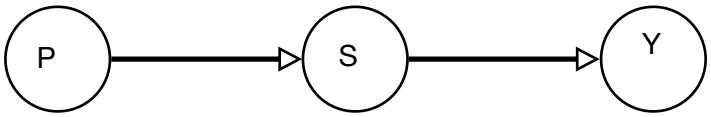
The present section provides a justified derivation of the model based on the work of Computer Scientist Judea Pearl who introduced *do-calculus*, a formal theory for structuring the a reasoning in a manner that allows the derivation from observational data of results comparable to those that would have been obtained in an experimental context (*see* Pearl 1995, 2010, 2012). The main logic of *do-calculus* consists in eliminating certain function from the model, and to replace them with arbitrary constants defined by the researcher. In this manner, it is possible to simplify a model to a subset of the original relations that it encompassed. This justifies, from the theoretical perspective, the use of structural equations, i.e. equations that specify the linkages between variables. However, in order to provide a more intuitive understanding, the model will be described, equivalently, using a Directed Acyclic Graph (or DAG).

DAGs can be described as sets of nodes and directed edges in which no cycles are permitted. To be sure, a DAG has already been introduced when Ltent Dirichlet Allocation was presented. However, in the present case the interest in the DAG is more than descriptive, it will be studied in order to identify *whether*, and in the next section *how*, it is possible to produce an estimate as comparable as possible to what would have been achieved had an experimental study been possible. Notably, such an experimental setting would be impossible under any respect: it would imply that it could be possible to force lobbyists to influence the policy process, and MEPs to vote according to different "strategies", or determinations (concepts that will be clarified in the following chapters).

In purely abstract, but reasonable, terms it is possible to assume that a certain *cause* that will be referred to as *strategy* determines with probability one, or *certainty*, the behaviour of the MEPs. Indeed, it may be objected that such a behaviour might not reflect reality, since people often choose without a strictly rational process. Three objections should be made to this criticism: 1) that modelling a perfectly rational behaviour is common in the discipline; 2) that the context under study is a politically relevant vote on

a legislative proposal, not a choice to be referred to ordinary life; 3) that it is possible to account, to a certain extent, for model uncertainty.

Therefore, it is not unreasonable to imagine that there exist certain probabilities $p_s \in P$, that determine the choice of a strategy from the set of all the possible strategies available $s \in S$; once the choice of a strategy has been made, it results in a precise outcome, a positive or negative vote $y \in Y$. A DAG would represent these relations as:



Moreover, it is useful to introduce the concept of *d-separation:*

Definition 2 ***d-separation*** (Pearl, 2010) A set S of nodes is said to block a path p if either (i) p contains at least one arrow-emitting node that is in S, or (ii) p contains at least one collision node that is outside S and has no descendant in S. If S blocks all paths from X to Y , it is said to "d-separate X and Y," and then, X and Y are independent given S, written $X \perp Y | S$.

On the basis of the criterion of d-separation it is possible to affirm that P and Y are *d-separated*, or that Y is conditionally independent from P when S is given $(Y \perp P \,|\, S)$. Hence, the causal query is equivalent to asking what would have been the value of Y if it was possible to experimentally change the strategy adopted by a given MEP (not possible in the physical world). Using the notation of *do-calculus,* the above problem may be expressed as:

$$P(Y = j \,|\, do(S = s)) \text{ with } j \in \{0,1\}$$

Notably, an intervention on S severs the link between the latter and P since the values of S are now chosen arbitrarily by the researcher in the context of the experiment. Therefore, the above probability can be interpreted as:

$$P(Y = j \,|\, do(S)) = P(Y = j \,|\, S = s) = P(j \,|\, s)$$

Therefore, provided that it is possible to obtain $P(j \,|\, s)$, or the probability of observing a certain category given a latent strategy, it should also be possible to assign a prior probability to each strategy, and to study the marginal the latter probability with

respect to the prior to obtain an estimate of the true value of the probability of the latent strategy under study. These problems will be developed in the following section.

6.3 Developing the model

Similarly to the logistic regression model it is possible to see the dependent variable as binomially distributed. In order to express the relationship between the qualitative ideas introducing this last model, and their quantitative translation, it is possible to think of the votes as instances of a binary variable, drawn from a discrete or "categorical" distribution, governed by an unknown parameter (theta).

$$y_i \sim Categorical(\theta)$$

Theta represents the probability of each strategy. In other words, it is possible to define $P(y_i \mid s_j) = \theta_j$, and express this quantity equivalently as (the equivalence holds because the equation below is effectively integrating out the p_s):

$$\theta_j = \sum_{s=1}^{2} p_s P(j \mid s) \quad \text{With } j \in \{0,1\}$$

Defined a set of strategies $\{1...s\}$, it has already been assumed that each should be associated with a probability p_s. Since the distribution of the latter probabilities is not given, it is necessary to assume an *a priori* distribution. Noting that p is a vector whose elements are positive and summing up to one (i.e. a simplex), since they represent probabilities for different strategies, it is reasonable to model it through a Dirichlet distribution (the initialisation is at four for each strategy, represents the idea that there is not a prior belief that one strategy should be more probable than the others): $p_s \sim Dirichlet([4,4])$, see 3.4).

More discursively, the model assumes that given a vote for MEP i, it was produced by a distribution parametrised according to a vector of strategies θ. This vector of strategies depends on the type of votes cast (in this case on whether the vote was favourable or not), and on the probability of the corresponding strategy that is assumed *a priori*. As such, all MEPs are considered equal in the present case (the same vector of prior probabilities is used to study the whole parliamentary assembly). In case of repeated observations it could be possible to assign different priors to each MEP.

However, the logistic regression has shown that different groups might have adopted different strategies depending on their political alignment (with groups on the left less likely to vote positively). These results fit uneasily with the assumptions of the current model. For these reasons, the model will be run both at the Parliamentary Assembly-level and at the political group-level.

What remains undefined, are therefore the strategies themselves, the causal structure underpinning the present analyses. Note that this structure should require a probabilistic interpretation, as it has been presented as $P(j \mid s)$. The probabilistic interpretation is the following: a certain sequence of vote j is observed with probability one when strategy s is present. This is a translation of the perfectly rational behaviour defended in 6.2. Whenever a certain strategy s is present the votes are assumed to follow a certain pattern with probability one.

The *career-seeking* strategy has been defined as the probability of voting favourably when the *urban rural* variable is above five and the national election system of the MEP is open. These two conditions should be met simultaneously. Moreover, in order to avoid confounding factors, it should be required that the MEP is not a follower of neither Copa Cogeca nor the Vegetarian Union. In this scenario, the MEP may be reasonably thought to be casting a positive vote because of constituency-related considerations. As a consequence, it would be expected that when the set of conditions does not hold, nothing may be said on the MEP's behaviour. Indeed, if they follow an interest group, or vote positively in the context of a constituency not interested in rural matters, or interested in rural matters but without a significant weight on MEP's career, nothing prevents that the cause of the observed behaviour might be different from career-seeking in the strict sense.

Defining a *policy-seeking* strategy is more complex, since it is not straightforward to devise a suitable variable. Nonetheless, from the perspective of the present work it seems reasonable to define this strategy as the probability of voting favourably when the MEP is a follower of Copa Cogeca (or related organisation). This assumption is derived from the hypotheses investigated in the literature, and from the idea that Twitter is both a social and an *information* network (Myers *et al.,* 2014).It should be noted that this definition does not depend on the constituency or on committee belonging.

Note that each of these definitions is made in terms of the positive vote, implying that whenever a vote not consistent with the conditions for each strategy is cast, it is considered to be zero under the relevant strategy. In this sense, the definitions may be interpreted as a mapping from the observations to vectors of binary values representing the level of fit between those observations and what would have been observed, had the strategy oriented the behaviour of the agents. For instance, if an MEP had casted a favourable vote under strategy one, having a level of *urban rural* below five, the mapping to the vector representing the fit of his behaviour to the strategy should be zero. On the other hand, if the following MEP's data respect all the conditions, it is assigned one in the vector. Computationally, what $P(j \mid s)$ is producing is a mapping from [1, 1] to [0, 1] under the first strategy.

6.4 Accounting for model uncertainty

As noted by the thesis supervisor the first strategy might be liable to considerable sensitivity due to the presence of the *urban rural* variable. Indeed, it is reasonable to assume that a high score of this feature (≥ 5) will produce an overestimation for the probability of the first strategy since it does not account for the favourable votes of MEPs linked to constituencies not interested in rural matters ($urban\ rural\ \leq 5$). Put it differently, it is a relevant to ask whether the behaviour may be assumed to be the same *independently* from the constituency.

The above problem is not resolved by eliminating the feature from the model. Arguably, the causal query is better addressed in terms of *model uncertainty*, i.e. how the baseline definition of the strategies is expected to change in response to a change in the conditions, this change should be understood not as a *non presence* of the feature, but as a *different degree* of presence. Eliminating the element of *urban rural* would make the strategy non-interpretable since there would not remain any valuable link between the MEP and the prevalent interest of his constituency.

It is possible to account for these differences using a technique known as Bayesian Model Averaging (BMA), consisting in a weighted average of the two models, where the weights, calculated through Bayes theorem, represent the probability of observing the data given the model (i.e. how well each model is able to represent the data). It is possible to express the probability of a model M_k as:

$$P(M_k \mid Y) \propto P(Y \mid M_k)P(M_k)$$

By choosing $P(M_1) = P(M_2) = \frac{1}{2}$ it is possible to encode the belief that neither model is considered a priori better than the other. Therefore, the relevant quantity to be derived is $P(Y \mid M_k)$. Before tackling this problem, it is necessary to expand on how the model updates its belief in more formal terms. Considering that it is expected that $P(y_i \mid \theta) = \theta_j$ whenever y_i is consistent with strategy s_j.

$$P(Y \mid \theta) = \prod_{i=1}^{N} P(y_i \mid \theta) = \prod_{i=1}^{N} \prod_{j=1}^{k} \theta_j^{I(y_i = s_j)}$$

The above likelihood can be expressed in a simplified form expanding the first product and in accordance with the indicator function *I:*

$$P(Y \mid \theta) = \prod_{j=1}^{k} \theta_j^{n_j} \text{ with } n_j = \sum_{i}^{n} I(y_i = s_j)$$

Using Bayes theorem with a Dirichlet prior for θ it is possible to define the probability of the strategies as:

$$P(\theta \mid Y) \propto P(Y \mid \theta) P(\theta) \propto \prod_{j-1}^{k} \theta_j^{n_j} \prod_{j=1}^{k} \theta_j^{\alpha_j - 1} \propto \prod_{j=1}^{k} \theta_j^{n_j + \alpha_j - 1}$$

The above result is analogous to what has been shown in the case of the Latent Dirichlet allocation model, the Dirichlet distribution of the hidden strategies is updated, by application of Bayes' Theorem as: $P(\theta \mid Y) \sim Dirichlet(\alpha_j + n_j, \ldots, \alpha_k + n_k)$. This updated parameters will be referred to as .

Now it is possible to answer the problem of departure, i.e. the derivation of $P(Y \mid M_k)$. Indeed, this probability should be equal to the likelihood marginalised over the possible values of the prior distribution. The following formula is equivalent to what has been derived above, except for a normalising constant (now present since the proportionality symbol is not being used).

$$P(Y \mid M_k) = \frac{1}{B(\alpha)} \int \prod_{j=1}^{k} \theta^{n_j + \alpha_j - 1} d\theta$$

Recalling that $\theta \sim Dirichlet(\xi)$ with $\xi = \alpha_j + n_j, \ldots, \alpha_k + n_k$, and that a probability distribution should sum to one, the integral above can be solved leveraging on this property, in this manner it is possible to estimate how probable are the observations given a certain model:

$$P(Y \mid M_k) = \frac{1}{B(\alpha)} \left[\int \prod_{j=1}^{k} \theta^{n_j + \alpha_j - 1} d\theta = \frac{1}{B(\xi)} \int \prod_{j=1}^{k} \theta^{\xi - 1} d\theta = B(\xi) \right] = \frac{B(\xi)}{B(\alpha)}$$

6.5 Application

The output of the model, adjusted for the uncertainty as presented in 6.3, has been fitted by means of the Metropolis-Hastings algorithm, and it is a distribution of probabilities whose mean will be taken as the estimate for the parameter governing the distribution of the observations, hence as the estimate of the probability of a certain strategy.

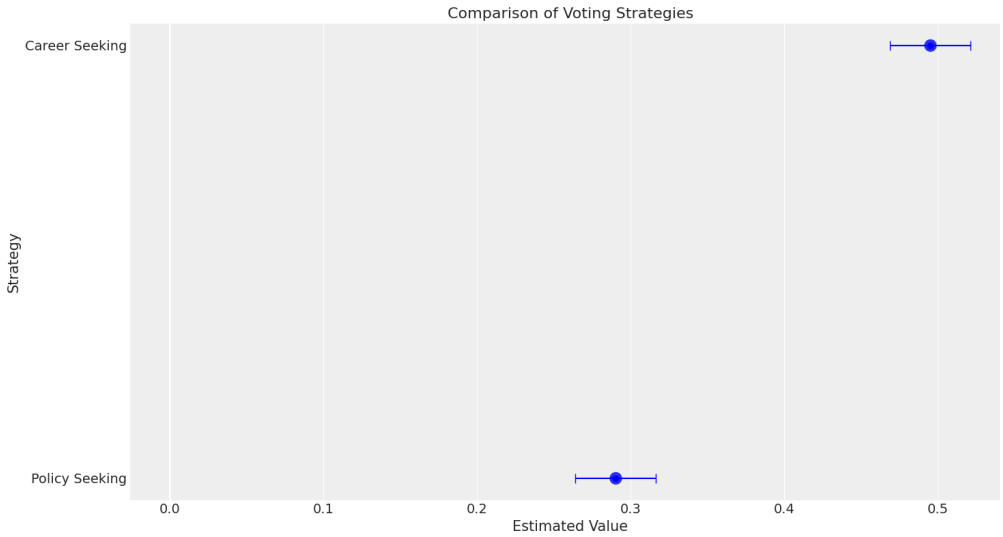
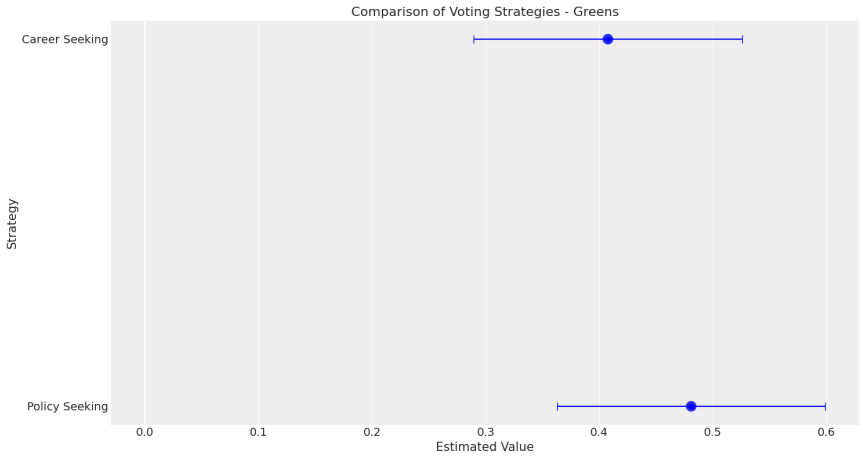


Fig. 5 Adjusted estimation of probabilities at the Assembly level. The Career Seeking strategy explains he observations more extensively.

The adjusted model confirms the thesis that the vote of the MEPs in this legislative proposal could be explained more by career seeking attitudes, than by career seeking on the basis of the strategies presented. However, the alteration of the first strategy in order to take into account of the potential uncertainty in he estimates, has revealed that, considering the whole parliamentary assembly, it is not possible to discern the career and policy seeking strategies whenever the costituency of the MEP is not interested in rural matters. In that case, the probability of an MEP voting favourably under each strategy is the same. Therefore, the optimistic estimate of 0.7 for the first

strategy, has been reduced to 0.5. However, since the second strategy has not been affected by the shift, the final results have not contradicted the thesis.

It should be noted, that the same conclusions do not necessarily hold if the analysis is brought to the party level. Indeed, while for all the parties the career seeking strategy is more probable than policy seeking, it is not so for the Greens and the S&D (from which the proposal under study originally came). These estimates should be taken with care since the standard deviations largely overlap in both cases.



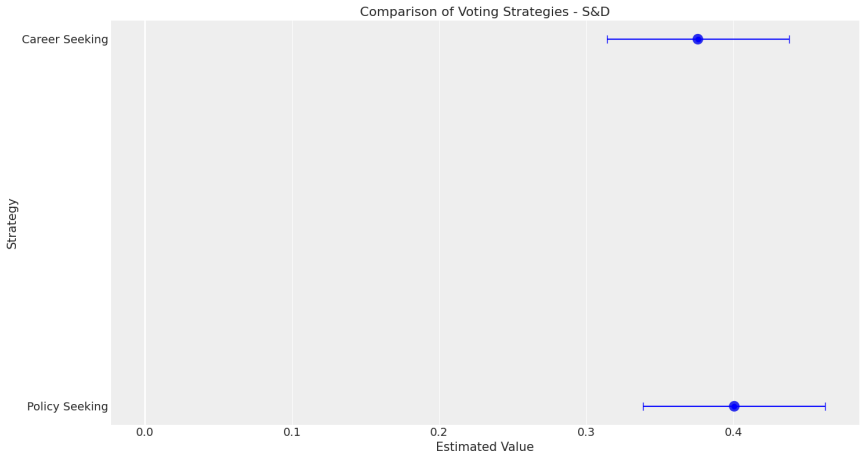


Fig. 6 and 7 Voting strategies for the Greens (above) and S&D (below).

Conclusions

The present thesis has investigated the legislative behaviour of MEPs relative to a law aimed at limiting the meaty denominations for vegan and vegetarian products. The research question was whether the behaviour of MEPs could be characterised as policy or career seeking, and the thesis that it could be inscribed in the latter category. In order to answer the question three models have been deployed: 1) through Latent Dirichlet Allocation it was possible to show that certain interest groups actively engaged to (possibly) influence the policy process; 2) through a multi-intercept logistic regression it was shown that the expected behaviour of the MEPs could not be decoupled from their belonging to a party and to a certain constituency; 3) Building on 1) and 2) a causal inference has been proposed in order to estimate the probability of two hidden strategies given the observed votes.

Following the above structure, each model has produced valuable insights that have been integrated in the successive model: from Latent Dirichlet Allocation the idea that not only Copa Cogeca, but also other stakeholders could be interested in a positive vote for this piece of legislation; from the logistic regression that an Assembly-level analysis might be insufficient; from the third model, instead, that the original definition of the strategies should have been revised and that the original interest from which the thesis departed could not be attained.

As regards the last point, as recalled in the introduction, the idea for this thesis stemmed from the persuasion that it could be possible to isolate the effect of interest groups on the policy process. These were optimistic hopes, not considering the available data. What could be answered satisfactorily was instead the current question, implying a broader understanding of the strategies as originally conceived. Indeed, voting according to a *policy seeking* strategy more than a *career seeking* one is not restricted to a pattern of favourable voting depending on the following of certain interest groups on social media, it might as well refer to a voting pattern that, not explained by career preoccupations, is more consistent with the idea of voting unfavourably when no apparent link between a lobby and an MEP can be found (see, for instance, the Greens).

Although restricted in scope, the present work contributes doubly to the current debate. Firstly, with the introduction of a causal model instead of an associational one (an approach not followed in the literature reviewed). Secondly, the originality of the subject which has not been explored elsewhere in this depth.

From the perspective of the utility of the results, it is argued that the present thesis offers some insights in the policy process. In particular, it is possible to

conclude, at least for the case at hand, that a general emphasis on constituency and national orientation has prevailed in the context of a politically divisive matter: an interpretation that confirms an intuitive understanding of EU politics. Less intuitive was the conclusion that the two strategies are not distinguishable when an MEP is linked to a constituency not interested in rural matters (at least when observing the Parliament as a whole). This result seems to be suggesting that for such MEPs there is less risk of political exposure, and therefore a zone of indifference that might be exploited by interested groups.

In conclusion, while perfectibility is always possible, the present work constitutes a first approximation for further developments, taking into account both the results that have been analysed in the thesis, and those that have been produced incidentally. Possible areas of further investigation include the reason why committee-belonging was not found significant, the other LDA results, or the extension of the present results integrating multiple observation per each MEP. At least for the denominations of meat-substitutes, it seems that the political willingness is determined more by pragmatic concerns than a programmatic approach, partly contradicting the centrality usually imputed to the agricultural lobbies.

Bibliography

Books and Journals

- Baratta, R. (2022). Institutions of EU law. Padova: CEDAM.

- Becker, G. S. (1983). A theory of competition among pressure groups for political influence. *The quarterly journal of Economics*. 98, pp. 371-400.

- Buchanan, J., M. and Tullock, G. (1962). *Logical Foundations of Constitutional Democracy* in The Collected Works of James M. Buchanan, vol: *3* The Calculus of Consent. Indianapolis: Liberty Fund.

- Frey, B. (1978). Politico-economic models and cycles. *Journal of Public Economics*. 9:2, pp. 203-220.

- Carreño, I., Dolle, T. (2018). Tofu Steaks? Developments on the Naming and Marketing of Plant-based Foods in the Aftermath of the TofuTown Judgement. *European Journal of Risk Regulation*. 9: 575-584. DOI: https://doi.org/doi:10.1017/err.2018.43.

- Carreño, I. (2022). France Bans "Meaty" Terms for Plant-Based Products: Will the European Union Follow?. *European Journal of Risk Regulation*. 13: 665-669. DOI: https://doi.org/10.1017/err.2022.22.

- Carrubba, C. et. al. (2006). Off the Record: Unrecorded Legislative Votes, Selection Bias and Roll-Call Vote Analysis. *British Journal of Political Science*, 36, pp. 691-704. DOI: https://doing.org/10.1017/S0007123406000366.

- Coate S. and Morris S. (1995). On the Form of Transfers to Special Interests. *The Journal of Political Economy*. 103:6, pp. 1210-1235. DOI: https://doi.org/10.1086/601449.

- Crew, M.A., Twight, C. (1990). On the efficiency of law: A public choice perspective. *Public Choice*. 66, pp. 15-36.

- Daniel, T. W., Obholzer, L., Hurka, S.(2019). Static and dynamic incentives for Twitter usage in the European Parliament. *Party Politics*. 25:6, pp. 771-781. DOI: 10.1177/1354068817747755.

- De Martini, E. et al. (2022). Would you buy vegan meatballs? The policy issues around vegan and meat-sounding labelling of plant-based meat alternatives. *Food Policy*. 111. DOI: https://doi.org/10.1016/j.foodpol.2022.102310.

- Fuentes, C., Fuentes, M. (2017). Making a market for alternatives: marketing devices and the qualification of a vegan milk substitute. *Journal of Marketing Management,* 33(7-8), pp. 529-555. DOI: 10.1080/0267257X.2017.1328456.

- Greene, D. and Cross, J. P. (2017). Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach. *Political Analysis.* 25:1, pp. 77-94. DOI: https://doi.org/10.1017/pan.2016.7.

- Hix, S. (2002). Parliamentary Behavior with Two Principals: Preferences, Parties, and Voting in the European Parliament. *American Journal of Political Science.* 46:3, pp. 688-698. DOI: https://doi.org/10.2307/3088408.

- Hix, S., Kreppel, A., Noury, A. (2003). The Party System in the European Parliament: Collusive or Competitive?. *Journal of Common Market Studies*. 41:2, pp. 309-31. DOI: https://doi.org/10.1111/1468-5965.00424.

- Hornung J., Bandelow N. C., Vogeler C. S. (2019). Social Identities in the policy process. *Policy Sciences*. 52:211-231. DOI: https://doi.org/10.1007/s11077-018-9340-6.

- Hornsey, M. J. (2008). Social Identity Theory and Self-categorization Theory: A Historical Review. *Social and Personality Psychology Compass*. 2(1), pp. 204-222. DOI: https://doi.org/10.1111/j.1751-9004.2007.00066.x.

- Ibenskas, R., Bunea. (2021). Legislators, organizations and ties: Understanding interest group recognition in the European Parliament. *European Journal of Political Research*. 60, pp. 560-582. DOI: 10.1111/1475-6765.12412.

- Koks, I. (2019). *Latent Dirichlet Allocation, explained and improved upon for applications in marketing intelligence*. Master Thesis. Delft University of Technology.

- Myers, S. A. *et al.* (2014). Information network or social network?: the structure of the Twitter follow graph. *WWW '14 Companion: Proceedings of the 23rd International Conference on World Wide Web*. DOI: https://doi.org/10.1145/2567948.2576939.

- Pearl, J. (1995). Causal Diagrams for Empirical Research. *Biometrika*. 82(4): pp. 669-710. DOI: https://doi.org/10.2307/2337329.

- Pearl, J. (2010).An introduction to Causal Inference. *The International Journal of Biostatistics*. 6(2). DOI: https://doi.org/10.2202/1557-4679.1203.

- Pearl, J. (2012). The do-calculus revisited. *Proceedings of the Twenty-Eight Conference on Uncertainty in Artificial Intelligence*. pp. 1-3. DOI: https://dl.acm.org/doi/10.5555/3020652.3020654.

- Klüver, H., Spoon, Jae-Jae. (2015). Bringing Salience back in: explaining voting defection in the European Parliament. *Party Politics*. 21:4, pp. 553-564. DOI: 10.1177/1354068813487114

- Krehbiel, K. (1993). Where's the Party?. *British Journal of Political Science*. 23:2, pp. 235-266. DOI: https://doi.org/10.1017/S0007123400009741.

- Lähteenmäki, A. et al. (2021). Alternative Proteins and EU food law. *Food Control*. 130:1-11. DOI: https://doi.org/10.1016/j.foodcont.2021.108336.

- Leialohilani, A., de Boer, A. (2020). EU food legislation impacts innovation in the area of plant-based dairy alternatives. *Trends in Food Science & Technology.* pp. 262-267. DOI: https://doi.org/10.1016/j.tifs.2020.07.021.

- Lindenberg, S. (2001). Social rationality versus rational egoism. In J. H. Turner (Ed.), *Handbook of socio- logical theory,* pp. 635–668. New York: Kluwer Academic.

- Mahmadou, V. (2022). Allocating Reports in the European Parliament: How parties influence committee work. *EPRG Working Paper, No. 7*. University of Tampere.

- Marshall, D. (2015). Explaining Interest Group Interactions with Party Group Members in the European Parliament: Dominant Party Groups and Coalition Formation. *Journal of Common Market Studies*. 53:2, pp.311-329. DOI: 10.1111/jcms.12163.

- McElreath, R. (2020). *Statistical Rethinking. A Bayesian Course with Examples in R and STAN*. New York: Chapman and Hall/CRC. DOI: https://doi.org/10.1201/9780429029608.

- Noury, A. (2002). Ideology, Nationality and Euro-Parlamentarians. *European Union Politics*. 3:1, pp.38-58. DOI: https://doi.org/10.1177/1465116502003001003.

- Pasour, E. C., Jr. (1992).Economists and public policy: Chicago political economy versus conventional views. *Public Choice*. 74: pp. 153-167.

- Pisanello, D., Ferraris, L. (2018). Ban on Designating Plant Products as Dairy: Between Market Regulation and Over-Protection of the Consumer. *European Journal of Risk Regulation.* 9: 170-6. DOI: doi:10.1017/err.2018.4.

- Sochirca, N. (2028). The European Legal Framework on Vegan and Vegetarian Claims. *European Food and Feed Law Review.* 13:6, pp. 514-521.

- Vallone, S. and Lambin, E. F. (2023). Public policies and vested interests preserve the animal farming status quo at the expense of animal product analogs. *One Earth*. 6, pp, 1213-1226. DOI: https://doi.org/10.1016/j.oneear.2023.07.013.

- Vogeler, C. S., Hornung, J., Bendelow, N. C. (2020). Farm animal welfare policy making in the European Parliament - a social identity perspective on voting behaviour. *Journal of Environmental Policy & Planning*. Vol. 22:4, pp. 518-530. DOI: https://doi.org/10.1080/1523908X.2020.1778458.

- Weiler, J. H. H. (1999). The Constitution of Europe. 'Do the New Clothes have an Emperor' and Other Essays on European Integration. Cambridge: Cambridge University Press.

- Yordanova, N. (2009). The Rationale behind Committee Assignment in the European Parliament. Distributive, Informational and Partisan Perspectives. *European Union Politics*. 10:2. pp. 253-280. DOI: 10.1177/1465116509103377.

- Yoshinaka, A., Mcelroy, G., Bowler, S.(2010). The appointment of rapporteurs in the European Parliament. *Legislative Studies Quarterly*. 35:5, pp. 457-486.

- You, H., Y. (2014). *Three Essays on Lobbying*. Doctoral dissertation. Harvard University.

Primary sources

Legal sources, European and national, chronologically ordered

- Regulation of the European Parliament and the Council 853/2004/EC laying down specific hygiene rules for on the hygiene of foodstuff' (2004) *Official Journal* L139/55. Available at: https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2004:139:0055:0205:en:PDF#:~:text=This%20Regulation%20lays%20down%20specific,processed%20products%20of%20animal%20origin. (Accessed 15/5/2023).

- Regulation of the European Parliament and the Council 1169/2011/EU on the provision of food information to consumers' (2011) *Official Journal* L304/18. Available at: https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2011:304:0018:0063:it:PDF. (Accessed 15/5/2023).

- Regulation of the European Parliament and the Council 1308/2013/EU establishing a common organisation of markets in agricultural products' (2013) *Official Journal* L347/671. Available at: https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:347:0671:0854:it:PDF. (Accessed 15/5/2023).

- Directive of the European Parliament and of the Council 2015/1535/EU laying down a procedure for the provision of information in the field of technical regulations and of rules on Information Society services' (2015) *Official Journal* L241/58. Available at: https://eur-lex.europa.eu/legal-content/IT/TXT/?uri=celex%3A32015L1535. (Accessed 27/5/2023).

- Judgement of the Court of 14/6/2017, *Verband Sozialer Wettbewerb eV v TofuTown.com GmbH,* C-422/16, EU:C:2017:458. Available at: https://eur-lex.europa.eu/legal-content/en/ALL/?uri=CELEX:62016CJ0422. (Accessed 21/5/2023).

- France. Assemblée national et Sénat (2018). *Loi pour l'équilibre des relations commerciales dans le secteur agricole et alimentaire et une alimentation saine, durable et accessible à tous*, No.2018-938, 30/10/2018. Available at: https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000037547946/. (Accessed 1/6/2023).

- France. Ministère de l' Économie, des Finances et de la Souveraineté Industrielle et Numérique (2022). *Décret relatif à l'utilisation de certaines dénominations employées pour désigner des denrées comportant des protéines végétales,* No. 2022-947, 29/7/2022.Available from: https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000045978360. (Accessed 12/6/2023).

Questions and answers to the Commission, paired sequentially, chronologically ordered

- Sommer, R. (2016). Deception with vegetarian and vegan foods. Question for written answer to the Commission*,* E-003771/2016. Available at: https://www.europarl.europa.eu/doceo/document/E-8-2016-003771_EN.html. (Accessed 27/5/2023).

- Mr. Andriukaitis (on behalf of the Commission). (2016). Answer to written question, E-003771/2016. Available at: https://www.europarl.europa.eu/doceo/document/E-8-2016-003771-ASW_EN.html. (Accessed 27/5/2023).

- De Castro, P., La Via, G. (2017). Vegan food products with misleading sales denominations. Question for written answer to the Commission, E-004044/2017. Available at: https://www.europarl.europa.eu/doceo/document/E-8-2017-004044_EN.html. (Accessed 27/5/2023).

- Mr. Andriukaitis. (2017). Joint answer on behalf of the Commission Written questions: E-004044/17, E-003755/17. Available at: https://

www.europarl.europa.eu/doceo/document/E-8-2017-003755-ASW_EN.html. (Accessed/23/5/2023).

- Hazekamp, A. (2018). French legislation concerning the naming of meat substitutes. Question for written answer to the Commission, E-002791/2018. Available at: https://www.europarl.europa.eu/doceo/document/E-8-2018-002791_EN.html. (Accessed 27/5/2023).

- Mr. Andriukaitis (on behalf of the Commission). (2018). Answer to written question, E-002791/2018. Available at: https://www.europarl.europa.eu/doceo/document/E-8-2018-002791-ASW_EN.html. (Accessed 27/5/2023).

Official documents, institutional and private entities, chronologically ordered

- Commission of the European Communities. (2007). *White paper on a strategy for Europe on nutrition, overweight, and obesity related health issues.* COM(2007) 279 final, 30/5/2007. Available at: https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2007:0279:FIN:EN:PDF. (Accessed 12/6/2023)

- European Commission. (2018). *Detailed information on the follow-up by the Commision to REFIT platform opinions.* Annual Burden Survey. OIB: Brussels. Available from: https://op.europa.eu/en/publication-detail/-/publication/1c578fe8-bd73-11e9-9d01-01aa75ed71a1#. (Accessed 12/6/2023).

- European Parliament: Committee on Agriculture and Rural Developement (2019). ****I Report on the proposal for a regulation of the European Parliament and of the Council amending Regulations (EU) No 1308/2013.* A8-0198/2019. Available at: https://www.europarl.europa.eu/doceo/document/A-8-2019-0198_EN.html. (Accessed 20/6/2023).

- BEUC. (2020). *'Veggie burgers' to remain 'burgers' thanks to EU Parliament vote.* Available from: https://www.beuc.eu/news/veggie-burgers-remain-burgers-thanks-eu-parliament-vote. (Accessed 23/5/2023).

- European Parliament (2021). CAP Amending Regulation (CMO) Amending regulations on the CMO for agricultural products, quality schemes and measures for remote regions. *Eu Legislation in Progress*, Briefing, Authored by: Beata Rojek,

PE 642.234. Available at: https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2019)642234. (Accessed 20/6/2023).

- European Vegetarian Union, FoodDrinkEurope and EuroCommerce. (2021). *Joint statement on vegan vegetarian definitions.* 16/9/2021. Available from: https://www.fooddrinkeurope.eu/resource/joint-statement-vegan-and-vegetarian-definitions/. (Accessed 20/6/2023).

Newspapers

- Audino, A. (2019). What's in a name: Veggie disc replaces veggie burger in EU food labels. *Slow Food.* 10 April. Available at: https://www.slowfood.com/whats-in-a-name-veggie-disc-replaces-veggie-burger-in-eu-food-labels/. (Accessed 14/4/2023)

- Fortuna, G. (2019). MEPs rubber-stamp first portion of next CAP, shed spotlight on wine and 'real' steak. *Euractiv.* 2 April. Available at: https://www.euractiv.com/section/agriculture-food/news/meps-rubber-stamp-first-portion-of-next-cap-shed-spotlight-on-wine-and-real-steak/. (Accessed 14/4/2023)

- Stone, J. (2019). Veggie burgers renamed 'veggie discs' under proposed new EU food labelling rules. *Independent.* 4 April. Available at: https://www.independent.co.uk/news/uk/politics/veggie-burgers-sausages-eu-steak-meat-industry-food-a8854961.html. (Accessed 14/4/2023).

- Liu, S. (2019). Dirichlet distribution. Motivating LDA. *Medium*. 6 January. Available at: https://towardsdatascience.com/dirichlet-distribution-a82ab942a879#:~:text=The%20Dirichlet%20distribution%20Dir(α,prior%20distributions%20in%20Bayesian%20statistics. (Accessed 3/6/2024).

APPENDIX I

LDA RESULTS

Index	Keywords	Interpretation
1	rural area woman cap cooperative gender opportunity young support policy	Support to women in agriculture
2	binding fertilisation spain population october programme situation eating sustainable	Amendments of Spanish law regarding fertilisers
3	eu cage sow cost egg production meat scenario sector impact	Egg and meat production
4	iyph explain care prof fruit edition simply simultaneously represent eastern	Representation of Eastern European producers of fruit
5	denomination cecinestpasunsteak plantbased know meat marketing imitation renew meatdenominations cultural	Misleading marketing of meat-substitutes
6	read smart join raise locally important virtual concerning review list	Virtual platform for engaging local communities
7	eaagrifood food worker trade organisation geopa employer war ukraine social	Impact of Ukraine war on agri food sector
8	used affecting poultry range survey unable farm whilst country pedro	Challenges in poultry farming
9	carbon emission energy biofuels forestry fuel farming removal commission proposal	Commission proposal on biofuels
10	aquaculture fish fishery organic regulation european feed plan action welfare	Fisheries policy

Index	Keywords	Interpretation
11	organic produc ed ent ce mar er protein ensur regulation	European protein strategy
12	fertiliser uan dumptheantidumping situation investigation import faced duty stakeholder fr	Use of fertilisers
13	contribution ass targetsarenotsolutions range agri sign summit economic shift president	Critique of target-based obligations on farmers
14	la ng le im st th ro sit new en	Not interpretable
15	addressed pdo partner coordinate opposed employer let security afternoon hope	Protection designation of origin
16	eu farmer food sector european production sustainable agricultural agriculture product	Sustainable European agriculture
17	animal welfare transport rabbit slaughterhouse legislation health breeding export like	Animal welfare during transports
18	tune effort com providing physical come securing northern wish enormous	Not clear.
19	sign jan measure inform largescale sustainably stronger quality low order	Scalability of sustainability practices
20	soil forester manager land strategy local essential forestry legislation law	Forest soil management
21	report policy hemp representative future letter return latest daily industrial	Hemp cultivation support.
22	award innovation womenfarmers woman project finalist farmer winner ruralwomeneu eufarmfork	Support to women in agriculture

Index	Keywords	Interpretation
23	agricoopforum harvest china affair chain ai gi addressed collecting small	Not clear.
24	forest eu strategy management owner wood need forestry european sfm	Forest strategy
25	honey beekeeper traceability party directive bee stanislav origin originating ec	Traceability of honey production
26	connection safety behaviour hiring living extra clarification sufficiently progress specialty	Promotion of correct behaviour on workplace
27	plant rice crop substance breeding ppp variety protection technique nbt	Rice crops protection
28	farmer today sector covid president key work eu importance impact	Impact of Covid on food sector
29	intervention objective national specific set type farmer support article condition	National support plans for farmers
30	eu cage sow cost egg production meat scenario sector impact	Use of cages for poultry and chicken.