Looking for a Better Shape: Societal Demand and Scientific Research Supply on Obesity

Lorenzo Cassi[1], Ismael Rafols[2], Pierre Sautier[3] and Elisabeth de Turckheim[4]

[1] *lorenzo.cassi@uni-paris1.fr*
Observatoire des Sciences et Techniques (HCERES-OST) and CES University of Paris 1 Pantheon-Sorbonne, Paris (France)

[2] *i.rafols@ingenio.upv.es*
Ingenio (CSIC-UPV), Universitat Politècnica de València, València (Spain),
SPRU (Science and Technology Policy Research), University of Sussex, Brighton (UK),
and Observatoire des Sciences et Techniques (HCERES-OST), Paris (France)

[3] *pierre.sautier@obs-ost.fr*
Observatoire des Sciences et Techniques (HCERES-OST), Paris (France), and Ingenio (CSIC-UPV), Universitat Politècnica de València, València (Spain)

[4] *elisabeth.deturckheim@obs-ost.fr*
Observatoire des Sciences et Techniques (HCERES-OST), Paris (France), and INRA, Délégation à l'évaluation, Paris (France)

Abstract
As science policy shifts towards an increasing emphasis in societal problems or grand challenges, new scientometric tools are required to inform decision-makers. However, while traditional bibliometrics could focus on the knowledge production side (the science supply), grand challenges also demand to investigate the articulation of societal needs. In this paper, we present an exploratory investigation of the grand challenge of obesity -an emerging health problem with enormous social costs. We illustrate a potential approach, showing: (a) how scientific publication can be used to describe existing science supply by using topic modelling based on publication abstracts; (b) how question records in the French parliaments can be used as an instance of social demand; and (c) how the comparison between the two may show (mis)alignments between societal concerns and scientific outputs.

Conference Topic
Science policy and research assessment

Introduction
Tackling complex global problems or grand challenges – such as climate change, food security, poverty reduction, risk of global pandemics – requires not only to increase R&D expenditure, but also the exploration and eventually the coordination of a variety of stakeholders with different areas of expertise and pursuing diverse research avenues. Typically these challenges benefit from the understanding of the physical and biological phenomena underlying a challenge (e.g. the virus and its genes), but also demand an understanding of the environmental and social contexts in which they occur, and the policy networks and instruments available in those contexts (Ely, Van Zwanenberg & Stirling, 2014).
Science policy funding schemes for societal problems or grand challenges seek to align science supply with social problems or needs. Although science is conducted in conditions of incomplete knowledge, it is well documented that certain particular research options are much

better aligned to certain outcomes (Sarewitz, 1996, pp. 31–49). It is, for example, very unlikely that astrophysics be useful for improving health care in malaria. Historically, several lines of inquiry in science policy have explored the alignment between research options and social outcomes, namely related to priority-setting and evaluation of research, but also to broader considerations related to the "supply" and "demand" of policy-relevant science. For this reason, a suitable interpretation of the alignment issue should be based on our understanding of the current state of the science (the *supply*) and what is required to achieve social goals (the *demand*) (Sarewitz & Pielke, 2007). The "demand" side must therefore consider not only the plurality of outcomes, but also various ways of articulating specific science or technology -driven pathways for achieving them. This in turn can refer to a process of public deliberation whereby different outcome preferences or divergent underlying values are made explicit by stakeholders. Similarly, the "supply" side is not just about how much "high-risk, high-return" research should be undertaken, but also about what type of outcomes are more or less *likely* to result from a given line of research. In this article, one the one hand, we apply the concept of *research landscape* (Wallace & Rafols, 2014) in order to map the scientific research on obesity.

On the other hand, we symmetrically map one of the interpretations (representations) of social needs (demand) on obesity. The supply-demand schema can be represented as in Figure 1. Here societal demand and scientific supply are not related directly in one single way. Instead they can relate via a variety of interpretation/representations of the "obesity" social needs. These representations shape science policy and affect actions that may reconcile supply and demand.

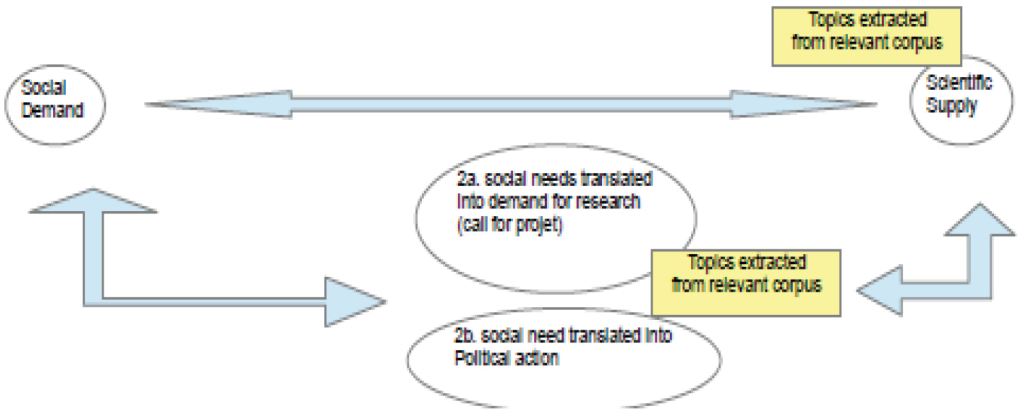

Figure 1. Social demand – scientific supply and political discourse as an example of intermediary representation.

In this paper we investigate the alignment (or lack thereof) between science supply and social demand by mapping, first, the scientific supply via the research landscape of obesity as defined by topic modelling of publication abstracts, and, second, social demand according to political discourse in the French parliaments. These maps of both supply and demand are specific and partial representations used in this preliminary and exploratory study -- other, complementary representations would be possible. For example, supply could be represented by a topic modelling of grants abstracts (as Talley et al., 2011 did for the US National Institutes of Health). And demand could be mapped using newspaper articles, among many other sources.

Data and Methods

Data

In order to define the relevant corpus for obesity, we follow a two-step method. First, we retrieve all publications with indexed MeSH term matching the search *obes** in MEDLINE/PubMed during the 2002-2013 period. This search was performed on October 16, 2014 and it returned 87,315 records.

Then, we launched *medlineR*, a routine based on the R language that allows the user to match data from Medline/PubMed with records indexed in the ISI Web of Science (WoS) database (Rotolo & Leydesdorff, 2015). The routine identified 71,055 WoS records (WoS core collections), with 'article' or 'review' as document types.

Second, we used Leiden's classification system to identify clusters of publications related to obesity. The classification system is constructed at the level of individual publications and clustering is based on direct citations (Waltman & van Eck, 2012) for the period 2000-2013. Obesity publications appear in 4,718 micro-clusters (in which at least one publication is tagged obesity), out of 32,466 micro-clusters for the whole WoS corpus. All the publications from clusters with at least 25% of publications tagged as 'obesity' were considered to be relevant for the study. This threshold of 25% is arbitrary and exploratory. Further explorations will use a lower threshold to test the robustness of this choice. The obesity corpus thus obtained contains 54,424 publications.

Topic modelling

Topic modelling provides a suite of algorithms to discover hidden thematic structure in large collections of texts. A topic model takes a collection of texts as input and it discovers a set of topics (recurring themes that are discussed in the collection) and the degree to which each document exhibits those topics.

Latent Dirichlet Allocation (LDA) is the simplest topic model. The intuition behind LDA is that documents exhibit multiple topics. LDA is a statistical model of document collections that tries to capture this intuition. It is most easily described by its generative process, the assumed random process. A topic is defined as a distribution over a pre-defined vocabulary. Moreover, it is assumed that the topics are specified before data have been generated (technically, the model assumes that the topics are generated first, before the documents). Now for each document in the collection, we generate the words in a two-stage process:

1. Randomly choose a distribution over topics.
2. For each word in the document
   (a) Randomly choose a topic from the distribution over topics in step #1.
   (b) Randomly choose a word from the corresponding distribution over the vocabulary.

This statistical model reflects the idea that each document contains multiple topics. Each document exhibits the topics with different proportion (step #1); each word in each document is drawn from one of the topics (step#2b), where the selected topic is chosen from the per-document distribution over topics (step #2a).

The goal of topic modelling is to automatically identify the topics from a collection of documents. The documents themselves are observed, while the topic structure (the topics, per-document topic distributions and the per-document per-word topic assignments) is a hidden structure.

Results on Science Supply

For this study, we fitted a 100-topic model to the 54,424 publications of the obesity corpus. We perform LDA with the R package “topicmodels” and visualize the output using LDAvis.

Figure 2 shows a map of these 100 topics. Topics are located close to one another if they are similar in terms of distributions of the words belonging to the selected dictionary. The measure of topic similarity is the matrix of Jensen-Shannon divergences between topics, considered as distributions over words, into two-dimensional coordinates and is represented in a 2d space through multi-dimensional scaling (i.e., principal coordinates analysis).
In addition, a clustering technique is used to cluster topics into research areas. We applied k-means clustering to the topics as a function of their two-dimensional locations in the global topic view with k=10. Labels are assigned to clusters. These labels are obtained by extracting the most relevant terms for each cluster of topics, where the term distribution of a cluster of topics is defined as the weighted average of the term distributions of the individual topics in the cluster.

Figure 2. Map of topics of publications on obesity (2003-2013).

Results on the Societal Demand

The same approach has been used to map the social demand. In order to define one possible interpretation, we refer to the questions that the members of the French Parliament (i.e. Assemblée Nationale or Senate) can ask to the government. Deputies and senators publicly question the members of the Government in different ways. The question can be asked during a Parliament session to the government or be written and a Parliament session is not necessary and addressed to one of the ministers. We retrieved this information and built up two datasets.

First, we selected all the questions asked by the Senators where the word *obes** was reported in the public database - with records from 1985 on - which is now available. We got 242 questions from 1992 - year of the first occurrence of 'obesity' in these questions - to 2014. Second, we collected oral and written questions asked by members of the Assemblée Nationale in the last three legislatures, getting: 422 questions (2002-2007), 870 (2007-2012) and 380 (2012 – 2014). The output of the 10-topic model is shown below for the Senate questions.

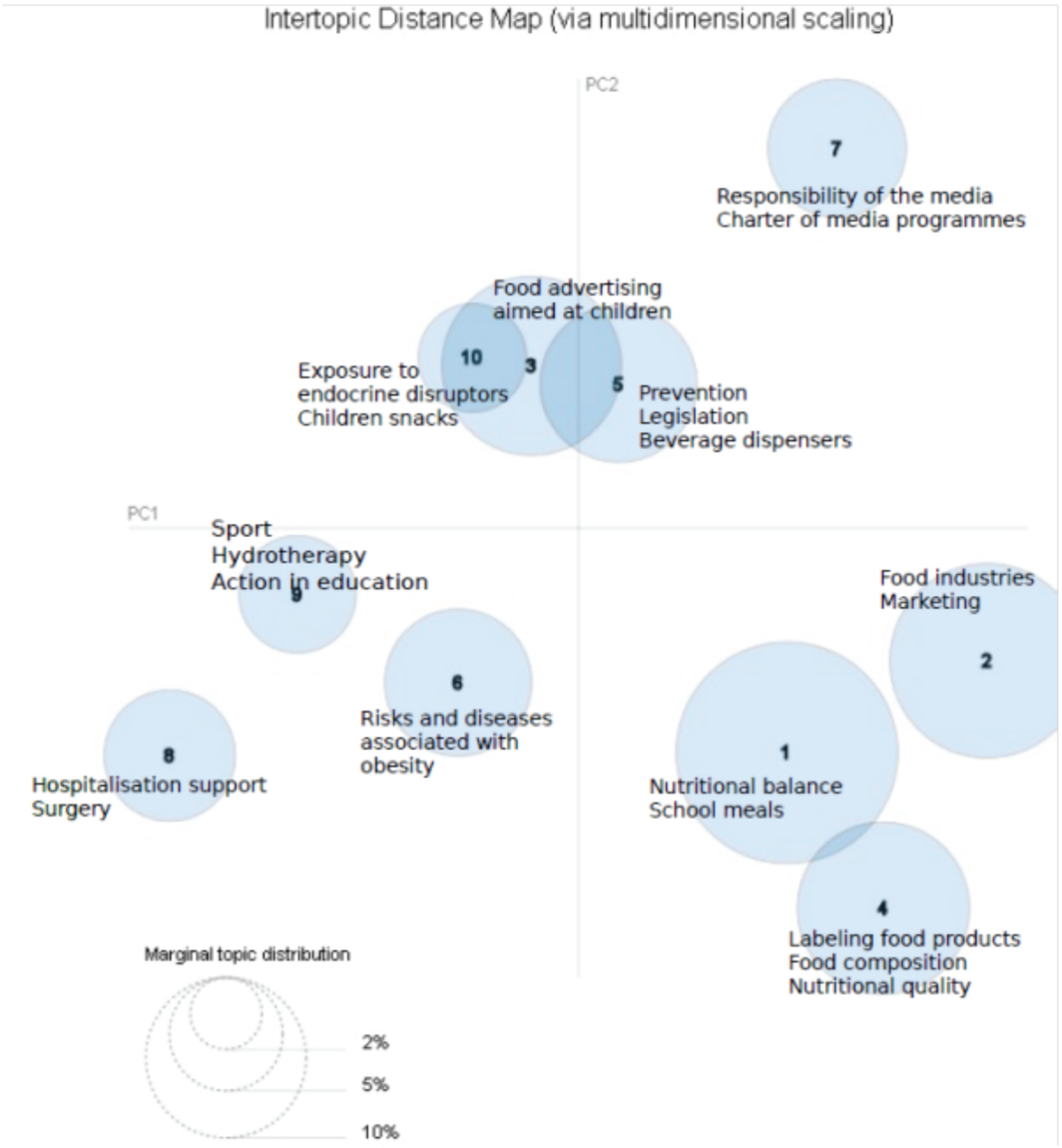

Figure 3. Map of topics for questions in the French Senate (1992-2014).

Discussion

In the centre of the Figure 2, we have a cluster of topics concerning *Physical activity, psychology and quality of life*, then turning around clockwise we find *Food consumption and diet* and then two clusters concerning mainly topics linked to biology research and further four clusters related to medical and surgery issues. The clusters of topics identified in the research landscape are mainly concerning medical and biological issues and only two clusters seem to deal with social and behavioural determinants of obesity, respectively *Obesity & ethnicity* and *Food consumption and diet*. The political discourse (Figure 3) seems to be organised around topics different from the research landscape. Among *the ten topics defined*

three main groups are reasonably identified. The first one, on the top part of the graph (i.e., topics number 3, 5, 7 and 10), is concerned mainly with children nutrition and the role of media as in advertising. A second group of topics, on the bottom right of the graph (i.e., topics number 1, 2 and 4), deals with food industry, marketing, and labelling issues. Finally, a third group, at the bottom left (i.e., topics number 6, 8 and 9) is concerned by medical and surgery issues. Only three out of ten topics of political discourse seem to find a counterpart in the research landscape. A preliminary analysis therefore suggests that, while research is concerned about the biophysical mechanisms that lead to obesity, many of the political questions are about the social mechanisms that favour obesity, such as advertisement, beverages, marketing, etc. This may suggest insufficient research regarding the social origin of obesity.

Acknowledgments

We would like to thank Tommaso Ciarli for suggesting to us the use of Parliament database as one of the possible representations of social needs. We thank Ludo Waltman for sharing the article level classification system.

References

Ely, A., Van Zwanenberg, P., & Stirling, A. (2014). Broadening out and opening up technology assessment: Approaches to enhance international development, co-ordination and democratisation. *Research Policy*, *43*(3), 505–518. doi:10.1016/j.respol.2013.09.004

Rotolo, D., & Leydesdorff, L. (2014). Matching MEDLINE/PubMed data with Web of Science (WoS): A routine in R language. *Journal of the Association for Information Science and Technology* (Forthcoming).

Sarewitz, D. (1996). *Frontiers of Illusion: Science, Technology and the Politics of Progress*. Philadelphia: Temple University Press.

Sarewitz, D., & Pielke, R. A. (2007). The neglected heart of science policy: reconciling supply of and demand for science. *Environmental Science & Policy*, *10*(1), 5–16.

Talley, E. M., Newman, D., Mimno, D., Herr, B. W., Wallach, H. M., Burns, G. A. P. C., & McCallum, A. (2011). Database of NIH grants using machine-learned categories and graphical clustering. *Nature Methods*, *8*(6), 443–444. doi:10.1038/nmeth.1619

Wallace, M. L., & Rafols, I. (2014). Research portfolios in science policy: moving from financial returns to societal benefits. *http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2500396*.

Waltman, L., & van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, *63*(12), 2378–2392. doi:10.1002/asi.22748