



**HAL**  
open science

## Le grand débat national, une aide pour prendre des décisions locales?

Jacques Fize, Lucile Sautot, Martin Lentschat, Ludovic Journaux, Mohamed Hilal

► **To cite this version:**

Jacques Fize, Lucile Sautot, Martin Lentschat, Ludovic Journaux, Mohamed Hilal. Le grand débat national, une aide pour prendre des décisions locales?. Spatial Analysis and Geomatics (SAGEO) 2021, May 2021, La Rochelle, France. hal-03375360

**HAL Id: hal-03375360**

**<https://hal.science/hal-03375360>**

Submitted on 12 Oct 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

---

# Le grand débat national, une aide pour prendre des décisions locales?

Jacques Fize<sup>1</sup>, Lucile Sautot<sup>2</sup>, Martin Lentschat<sup>3</sup>,  
Ludovic Journaux<sup>4</sup>, Mohamed Hilal<sup>5</sup>

1. INSA Lyon - LIRIS - jacques.fize@insa-lyon.fr

2. AgroParisTech - UMR TETIS - lucile.sautot@agroparistech.fr

3. Univ. de Montpellier - UMR TETIS - martin.lentschat@umontpellier.fr

4. AgroSup Dijon - LIB - ludovic.journaux@agrosupdijon.fr

5. INRAE - UMR CESAER - mohamed.hilal@inrae.fr

---

*RÉSUMÉ.* Le Grand Débat National, décidé par Emmanuel Macron début 2019 pour répondre au mouvement social des Gilets Jaunes, a permis de collecter les contributions de citoyens sur la transition écologique via une plateforme en ligne. Dans cet article, nous exploitons le corpus constitué par ces contributions pour identifier des zones où les participants demandent le développement de pistes cyclables et d'équipements ferroviaires. Pour cela, nous avons créé un modèle de classification permettant d'identifier les contributions traitant de la thématique du transport et proposé une méthode d'extraction de motifs traduisant les propositions des contributeurs. À l'aide des codes postaux donnés par les contributeurs, nous présentons les motifs extraits sur plusieurs cartes.

*ABSTRACT.* The Great National Debate, decided by Emmanuel Macron at the beginning of 2019 to respond to the Yellow Vests social movement, allowed the collection of citizens' contributions on the ecological transition via an online platform. In this article, we use the corpus constituted by these contributions to identify areas where participants are asking for the development of bicycle paths and railway facilities. For this purpose, we have created a classification model to identify contributions dealing with the theme of transportation and proposed a method for extracting patterns that reflect the contributors' proposals. We then represented these patterns on maps, using the contributors' postal codes.

*MOTS-CLÉS :* Grand Débat National, Motifs, Transport

*KEYWORDS:* Grand Débat National, Patterns, Transport

---

## Introduction

En octobre 2018, une hausse du prix du carburant automobile déclenche en France un mouvement spontané de protestations sociales dénommé mouvement des "gilets jaunes", avec des blocages de routes, de ronds-points et de nombreuses manifestations. La première réunit 280 000 personnes le 17 novembre 2018 (acte 1) et surprend par son ampleur (Boyer *et al.*, 2020). Elle sera suivie par de multiples formes de mobilisations accompagnées de revendications sociales et politiques. En décembre 2018, face à ce mouvement, le gouvernement français saisit la Commission nationale du débat public (CNDP<sup>1</sup>) afin de lancer une consultation citoyenne sur divers sujets : le *Grand Débat National*. Cette consultation s'est officiellement tenue du 15 janvier au 15 mars 2019 autour de quatre grands thèmes suivant : transition écologique, fiscalité, démocratie et citoyenneté, organisation de l'État et des services publics. La consultation s'est déroulé sous différentes formes : réunions publiques locales, cahier de doléances en mairie, conférences avec des citoyens tirés au sort, contributions en lignes sur une plateforme dédiée, etc.

L'analyse et la synthèse des contributions en ligne posent plusieurs défis. Tout d'abord, elles sont nombreuses. Sur la plateforme en ligne, pour le questionnaire de la *Transition Écologique*, plus de 134 000 personnes ont rempli 12 champs de texte libre, soit un total de 1,6 millions de textes potentiels<sup>2</sup> à analyser. De plus, les textes recueillis sont de longueurs différentes et présentent une grande diversité (style, syntaxe, orthographe, grammaire, lexique). À la demande du gouvernement français, une première analyse des contributions en ligne est effectuée par l'entreprise *OpinionWay*<sup>3</sup> immédiatement après la clôture du débat (Opinion Way, 2019). Les synthèses proposées offrent une vue d'ensemble du contenu des contributions mais présentent plusieurs limitations. Tout d'abord, la méthodologie utilisée n'est pas complètement documentée (Bellet *et al.*, 2020), ce qui rend difficile sa reproductibilité et la comparaison avec d'autres travaux. Ensuite, selon les questions analysées, un nombre important de réponses (entre 10 % et 30 %) est écarté. Enfin, la dimension territoriale inhérente tant au mouvement social à l'origine de la consultation qu'aux thématiques de la consultation a été très peu abordée.

Par ailleurs, peu de travaux scientifiques ont été publiés sur l'analyse de ces contributions. Parmi ces communications, les auteurs de (Ploux *et al.*, 2020) ont proposé une analyse automatique sur la base de motifs et d'une comparaison du lexique employé dans différentes consultations incluant celles du Grand Débat National. Cet article, disponible sur HAL, n'a a priori pas encore été publié dans une revue ou ouvrage à comité de lecture. Nous pouvons également cité

---

1. <https://www.debatpublic.fr/>

2. Les contributeurs ne répondent pas nécessairement à l'ensemble des questions.

3. Les synthèses *OpinionWay* sont disponibles ici : <https://granddebat.fr/pages/syntheses-du-grand-debat>

(Messer *et al.*, 2019), qui propose un approche qualitative des contributions relatives à la mobilité. Ce rapport, très riche sur la description des propositions formulées par les contributeurs, ne dit rien sur la dimension spatiale. En réponse à cela, nous proposons un traitement reproductible des données et une analyse des résultats tenant compte de la dimension géographique. Dans cet article, nous présentons les travaux suivants:

- **Focus sur les réponses de la "transition écologique"** et sur une sous-thématique cruciale : **les transports**.
- **Création d'un modèle de classification** pour identifier les réponses parlant de cette sous-thématique.
- **L'extraction de motifs** dans les données traduisant les propositions émises par les contributeurs.
- **Analyse géographique des motifs** extraits au travers de deux objets : le train et les pistes cyclables.

Cet article s'organise de la manière suivante. La section 1, présente l'approche utilisée pour classifier les réponses et extraire les motifs d'opinions. La section 2, présente les motifs extraits selon leur fréquence et leur localisation géographique. Enfin, nous proposons plusieurs discussions en section 3 sur les résultats obtenus, avant de conclure en section 4.

## 1. Matériel et méthodes

Dans ces travaux, nous réalisons une analyse des opinions sur la question du transport dans les réponses de la *transition écologique*. Ce processus s'effectue en deux étapes : la première (*c.f.* Section 1.2) identifie les réponses mentionnant le transport ; la seconde (*c.f.* Section 1.3) en extrait les motifs en s'appuyant sur des règles d'extraction et plusieurs lexiques. Le processus est illustré dans la Figure 1.

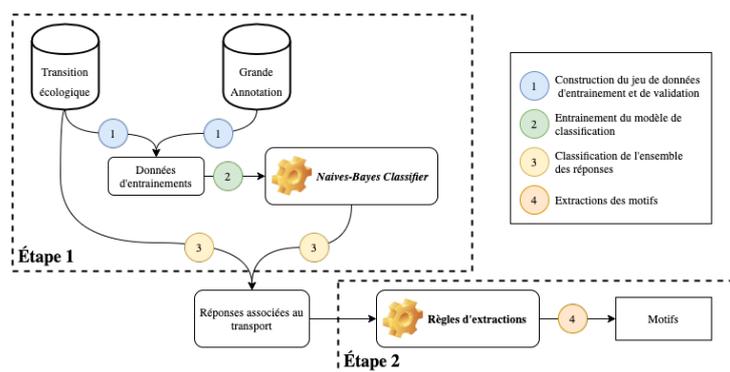


FIGURE 1 – Processus d'extractions des motifs d'opinions

Avant de présenter les méthodes proposées pour chaque étape, nous présentons les données utilisées dans l'ensemble du processus.

### 1.1. Données

Dans cet article, nous utilisons les contributions de la plateforme web du *Grand Débat National*<sup>4</sup>. La plateforme proposait deux modalités de contribution : libre ou QCM. Seules les contributions libres sont exploitées dans cet article. Plus particulièrement, nous avons uniquement retenu les contributions au volet portant sur la *Transition Écologique*.

Concernant l'analyse géographique, nous exploitons le code postal indiqué par chaque contributeur, sachant que 97 % de ses codes sont exploitables. Les coordonnées géographiques (latitude, longitude) sont extraites de la base officielle des codes postaux<sup>5</sup> et correspondent aux centroïdes.

Les réponses sont contextualisées avec des informations externes liées à l'offre de transport. Plus particulièrement, la présence des infrastructures<sup>6</sup> (les pistes cyclables, les gares, les voies ferrées) et l'organisation du territoire par les déplacements domicile-travail (base communale des aires urbaines définies par l'*INSEE*<sup>7</sup>).

### 1.2. Classification des réponses liées aux transports

Dans ce travail, nous nous concentrons sur une thématique centrale dans le cadre de la transition écologique : **le transport**. Pour cela, nous isolons les réponses liées à ce thème en utilisant les annotations réalisées, parallèlement à l'appel à contribution du grand débat, par le collectif nommée *Grande Annotation*<sup>8</sup>. Les réponses du Grand Débat sont annotées à l'aide d'une liste de *tags*. Au moment du téléchargement des données<sup>9</sup>, 280 321 réponses étaient annotées. Ce nombre d'annotations est insuffisant pour une analyse globale des réponses du Grand Débat. Toutefois, les annotations présentes sont suffisantes pour entraîner un modèle de classification. Notamment, si le seul objectif du modèle est d'indiquer si une réponse est associée ou pas à la thématique du transport. Pour établir cet ensemble, nous avons construit un jeu de données d'entraînement/validation composé de deux types de réponses, l'une évoquant le transport et l'autre non. L'identification des réponses annotées parlant de

---

4. <https://granddebat.fr/pages/donnees-ouvertes>

5. <https://www.data.gouv.fr/fr/datasets/base-officielle-des-codes-postaux/>

6. Données extraites sur *OpenStreetMap* <https://download.geofabrik.de/europe/france.html>

7. <https://www.insee.fr/fr/information/2115011>

8. Données accessible ici : <https://grandeannotation.fr>

9. 31 mai 2020

transport s’opère en sélectionnant celles ayant un des *tag* faisant partie de la liste présentée dans la Table 1.

Une fois les données d’entraînement et de validation extraites, nous avons testé différents algorithmes de classification. Parmi ceux-ci, nous comparons des approches classiques : SUPPORT VECTOR MACHINE (SVM)(Vapnik, 2013), NAIVE-BAYES (Manning *et al.*, 2008), KNN (Altman, 1992), DECISION TREE (Loh, 2011) et RANDOM FOREST (Breiman, 2001). De plus, nous ajoutons des approches plus récentes venant du *deep learning* : BI-LSTM et BERT (Devlin *et al.*, 2019).

TABLE 1 – Tags de la Grande Annotation associée à la thématique du transport

Transports doux ou moins nombreux	Pistes cyclables, sécurisées	Commerce de proximité, circuit-court
Déménagement ou changement travail	Obligation, incitation au transport collectif	Transport commun gratuit
Télétravail	Transport commun plus fiable, rapide	Covoiturage, auto-partage, stop
Pistes cyclables	Transport commun moins cher	Transport commun plus sûr
Transports en commun meilleurs	Télétravail, horaires flexibles	Parkings (vélo, auto) à proximité des gares
Transports en commun moins chers	Transport commun plus proche	Développer le réseau ferroviaire
Taxer les gros pollueurs	Transport commun plus fréquent	Ramassage scolaire
Vélo, trottinette (électrique)	Voiture électrique	Transports en commun (bus, tram)
Voiture plus propre	Marche à pied	Voiture à hydrogène
Voiture autonome	Train	Voiture hybride
A Déplaçons-nous moins	Transport fluvial	
Transport à la demande	Biocarburant (E85, ...)	

Pour la représentation des données, nous utilisons une représentation *bag of words* (sac de mots) pour les approches classiques. Pour résumer, une représentation *bag of words* encode un texte dans un vecteur  $v = \{v_0, \dots, v_N\}$  de dimension  $N$ , où  $N$  correspond au nombre de mots uniques présents. Chaque valeur  $v_i$  du vecteur correspond à la présence d’un terme ou au nombre d’occurrences de celui-ci dans la réponse. Ici, une valeur correspondent au nombre d’occurrences d’un mot pondéré par la formule TF-IDF (Gerard Salton, 1983). Avant de transformer les textes en *bag of words*, ceux-ci sont *lemmatisés*. La lemmatisation consiste à remplacer chaque mot par son *lemme*, c’est à dire la forme canonique d’un mot (*e.g* aimé  $\rightarrow$  aimer).

Dans la Table 2, nous présentons les performances obtenues par les différents algorithmes de classifications sélectionnées. Si les valeurs d’exactitude sont proches, ce sont les algorithmes NAIVE-BAYES et BERT qui présentent les scores les plus élevés. Une fois l’algorithme de classification ayant produit un modèle, nous utilisons celui-ci sur le reste des réponses du thème transition écologique. Les résultats de classification sont présentés dans la Table 3.

TABLE 2 – Résultat de classification selon différents algorithmes

Modèle	Paramètres	Exactitude
SVM	g=0.1 & kernel=rbf	0.86
NAIVE-BAYES	alpha = 1	<b>0.87</b>
SGD	loss=log & penalty = L1	0.86
KNN	n = 5 & p =2	0.76
DECISION TREE	criterion = gini	0.83
RANDOM FOREST	g=100 & kernel=rbf	0.84
BI-LSTM	32 LSTM cells	0.69
BERT	BERT multilingual base model (cased)	<b>0.87</b>

TABLE 3 – Pourcentage de réponses par question associées au transport

Question	Pourcentage
Quel est aujourd'hui pour vous le problème concret le plus important dans le domaine de l'environnement ?	0,59
Que faudrait-il faire selon vous pour apporter des réponses à ce problème ?	10,36
Si oui, de quelle manière votre vie quotidienne est-elle touchée par le changement climatique ?	1,05
Si oui, que faites-vous aujourd'hui pour protéger l'environnement et/ou que pourriez-vous faire ?	27,71
Qu'est-ce qui pourrait vous inciter à changer vos comportements comme par exemple mieux entretenir et régler votre chauffage, modifier votre manière de conduire ou renoncer à prendre votre véhicule pour de très petites distances ?	28,79
Quelles seraient pour vous les solutions les plus simples et les plus supportables sur un plan financier pour vous inciter à changer vos comportements ?	17,54
Si oui, que faudrait-il faire pour vous convaincre ou vous aider à changer de mode de chauffage ?	1,42
Si oui, que faudrait-il faire pour vous convaincre ou vous aider à utiliser ces solutions alternatives ?	34,08
Si non, quelles sont les solutions de mobilité alternatives que vous souhaiteriez pouvoir utiliser ?	59,51
Et qui doit selon vous se charger de vous proposer ce type de solutions alternatives ?	9,30
Que pourrait faire la France pour faire partager ses choix en matière d'environnement au niveau européen et international ?	3,75
Y a-t-il d'autres points sur la transition écologique sur lesquels vous souhaiteriez vous exprimer ?	5,40

### 1.3. Extraction de motifs associés à une opinion

Dans ces travaux, nous nous appuyons sur la définition d'une opinion de (Liu, Zhang, 2012). Une opinion  $o_i$  est définie selon le quintuplet  $\{o_e, o_p, o_c, o_a, o_d\}$ ;  $o_i \in O$ , où  $O$  correspond à l'ensemble des opinions présentes dans une unité de texte (phrase, paragraphe, etc.),  $o_e$  correspond à l'émetteur de l'opinion,  $o_p$  est le polarisateur,  $o_c$  la cible de l'opinion,  $o_a$  est l'attribut de la cible de l'opinion

et  $o_d$  est la date à laquelle l'opinion est émise. Par exemple, dans la phrase, "*Je trouve que le prix du billet de train est hors de prix.*", l'opinion  $o_1$  est {"Je", "hors de prix", "billet de train", "prix",  $\emptyset$  }. Dans ces travaux, nous ajoutons à ce quintuplet la (géo-)localisation de l'émetteur,  $o_l$ .

### 1.3.1. Un format d'opinion : la proposition

Il existe plusieurs manière d'exprimer une opinion. Parmi celles-ci, on peut trouver le constat et la proposition. Par exemple, la phrase "J'utilise les transports en communs" exprime un constat et "*Il faut augmenter le nombre de train*" exprime une proposition. Dans la première phrase, la cible sont "les transports en communs", le verbe utiliser est le polarisateur. Dans la seconde phrase, la proposition est d'"augmenter"(polarisateur) le "nombre"(attribut de la cible) de "train"(cible de l'opinion). Le problème du constat réside dans l'ambiguïté du sentiment qu'éprouve un acteur concernant la cible de l'opinion. Reprenant la phrase d'exemple, le contributeur utilise peut-être les transports en commun par contrainte financière et souhaite dans un avenir proche passer à un moyen de transport individuel. Étant moins ambiguë, nous préférons nous focaliser sur la forme d'opinion qu'est la proposition.

Dans ces travaux, reprenant le quintuplet de (Liu, Zhang, 2012). *L'émetteur*  $o_e$  correspond à l'auteur de la contribution. Le *polarisateur*  $o_p$  est porté par un verbe (modifier, développer, augmenter, etc.) ou adverbe (plus, moins, etc.) de changement. La cible de l'opinion  $o_c$  correspond à un objet appartenant à la catégorie de transport (moyen de locomotion, infrastructure, etc.). Si celle-ci peut être absente, l'attribut  $o_a$  de l'objet cible  $o_c$  est caractérisé par un nom (e.g. prix, fonctionnement, rapidité, etc.). La *date d'émission*  $o_d$  correspond à la date de publication sur la plateforme. La *localisation*  $o_l$  est indiquée par le code postal indiqué par le contributeur.

Dans ces travaux, nous avons choisi de traiter une thématique centrale de la transition écologique : le transport. Pour cette raison, les objets cibles  $o_c$ , les attributs  $o_a$  mais aussi les polarisateurs  $o_p$  sont délimités au sein des trois lexiques suivants :

1. **Le lexique du transport.** Ce lexique contient les termes désignant les moyens de locomotions (train, vélo, bus, voiture, etc.), les infrastructures (gare, voie de chemin de fer).
2. **Le lexique du changement.** Ce lexique contient des termes impliquant une modification (plus de, changer, augmenter, etc.).
3. **Le lexique d'attribut** associés aux termes de transports (prix, accessibilité, quantité, etc.).

L'extraction de motifs s'effectue de la manière suivante. Dans un premier temps, nous lemmatisons l'ensemble du document. Ensuite, nous identifions les termes des différents lexiques ainsi que leur position dans le document. En s'appuyant sur ces mêmes positions, nous identifions les termes qui appartiennent

à une même fenêtre dont la taille est définie par la constante  $t_{fen}$ . En conséquence, si un terme appartenant au lexique du changement et au lexique du transport apparaissent dans une même fenêtre, une opinion suivant un premier motif est considéré (polarisateur + cible). Si un terme du lexique des attributs se trouve entre les deux termes formant ce motifs, alors l'opinion suit un second motifs (polarisateur + attribut + cible).

Dans la section suivante, nous présentons les résultats de l'extraction de motifs obtenus en s'appuyant sur notre méthode et les données utilisées.

## 2. Résultats

### 2.1. Répartition des motifs obtenus

Dans la Figure 2, nous présentons la distribution des motifs selon le polarisateur et objet cible des opinions détectées.

Dans ces résultats, plusieurs tendances sont identifiables. La première concerne le développement des infrastructures tels que les pistes cyclables ou encore les réseaux de transport mais aussi de moyen de locomotions comme le covoiturage. Inversement, on observe une tendance négative envers les transports individuels et plus particulièrement la voiture, par exemple, associée très largement au verbe limiter, réduire et diminuer. Cependant, celle-ci est contrebalancée par une troisième tendance en lien avec le remplacement d'un véhicule souvent plus économique (diesel) et/ou plus écologique (hydrogène, électrique). On peut observer cette dernière tendance avec les verbes tels que acheter, changer et remplacer.

### 2.2. Répartition géographique des motifs obtenus

Notre analyse permet de mettre en évidence la localisation des contributeurs qui demandent un développement des pistes cyclables, et des contributeurs qui demandent un développement du réseau ferroviaire et/ou une augmentation de la fréquence des trains.

La Figure 3a présente une carte de la localisation des contributeurs qui demandent un développement des pistes cyclables au regard de la densité de pistes cyclables à l'échelle de la France métropolitaine. La Figure 3b présente une carte de la localisation des contributeurs qui demandent un développement du réseau ferroviaire et/ou une augmentation de la fréquence des trains au regard de la densité de rails à l'échelle de la France métropolitaine. Ces deux cartes permettent d'identifier approximativement les zones où des nouveaux équipements ou le développement des équipements existants sont requis par des contributeurs.

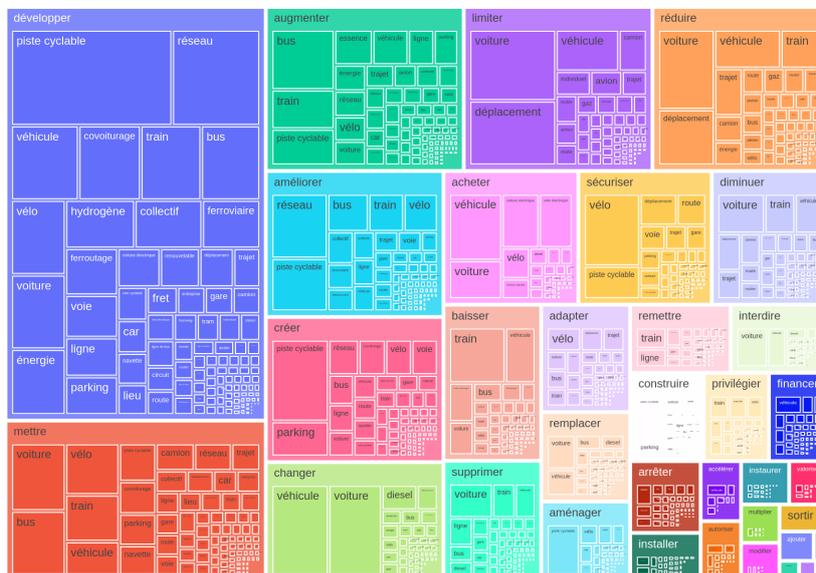


FIGURE 2 – Répartition des motifs selon leur polarisateur et objet ciblé

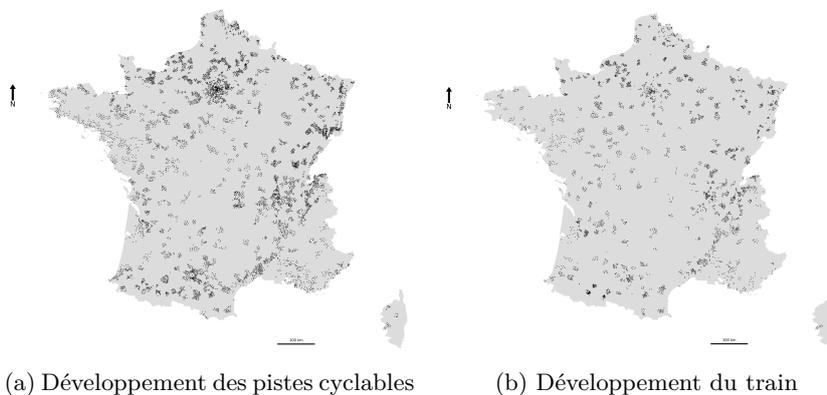


FIGURE 3 – Nombre de contributions selon les motifs sélectionnés

Une échelle d’analyse plus locale semble plus appropriée pour exploiter au mieux ces données. La conférence ayant lieu à La Rochelle, nous avons choisi, à titre d’exemple, de présenter les mêmes données à l’échelle de la Charente-Maritime (Figure 4 et Figure 5).

Les contributeurs étant localisés grâce à leurs codes postaux, une analyse fine à l’échelle de la commune ou de la communauté de commune est impossible. En revanche, il est possible de faire des recommandations à l’échelle des aires urbaines et des départements. Ainsi, en *Charente-Maritime*, il semble que le

développement de pistes cyclables soient demandé par des contributeurs dans les aires urbaines de *La Rochelle*, *Saintes* et *Saint Jean d'Angély*, ainsi que dans les communes situées entre *Saintes* et *Rochefort* et entre *Saintes* et *Cognac*. Concernant les demandes de développement du réseau ferroviaire et/ou une augmentation de la fréquence des trains, les besoins sont exprimés au centre de des aires urbaines de *Sainte* et de *La Rochelle*. On remarque également une demande dans les communes isolées au nord de *Saint Jean d'Angély*, entre les gares de *Saint Jean d'Angély* et de *Villeneuve-la-Comtesse*.

Les cartes produites par le croisement des contributions en ligne au Grand Débat Nationale, des aires urbaines et des équipements recensés par *OpenStreetMap* permettent donc de proposer des zones de demandes pour certains équipements spécifiques.

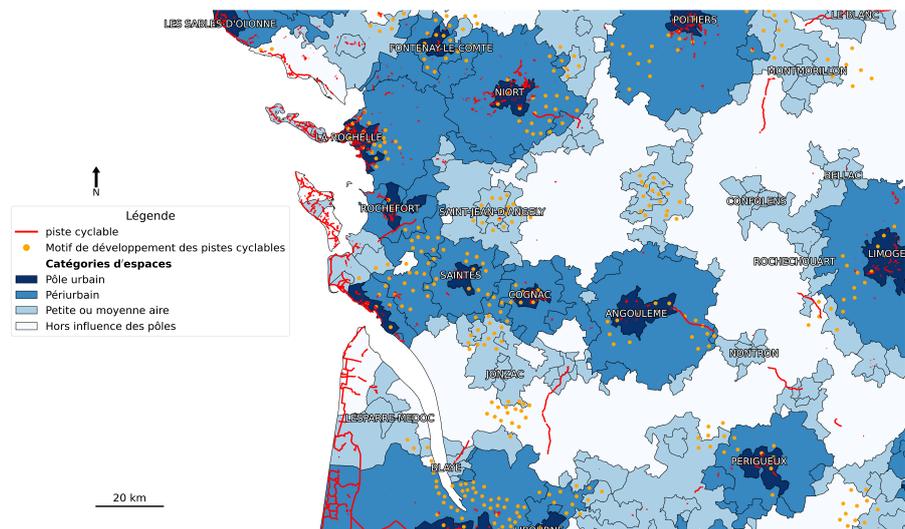


FIGURE 4 – Contributeurs demandant un développement des pistes cyclables en *Charente-Maritime* et en *Charente*

Les données et le code nécessaire à la réalisation des cartes sont mis à disposition en ligne.<sup>10</sup>

10. **Code.** [https://github.com/Make-the-Debat-Great-Again/map\\_generation](https://github.com/Make-the-Debat-Great-Again/map_generation); **Données.** <https://zenodo.org/record/4270184>

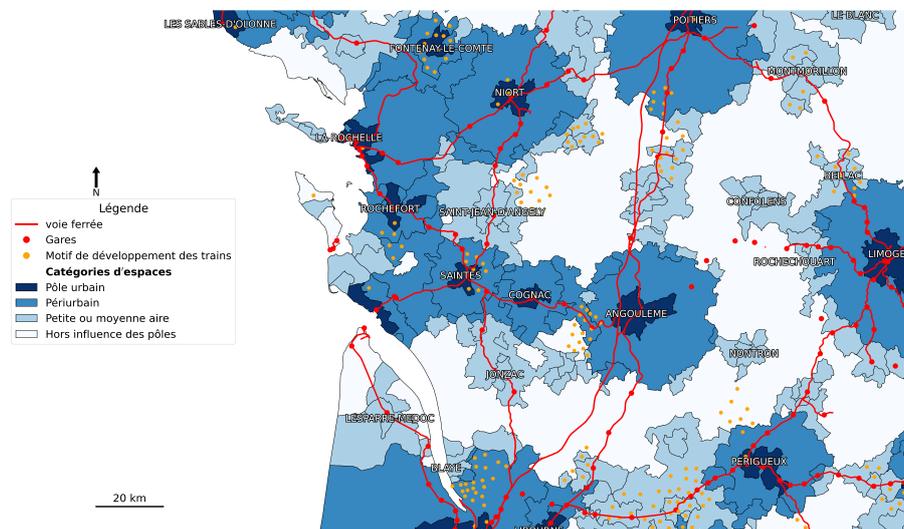


FIGURE 5 – Contributeurs demandant un développement du réseau ferroviaire et/ou une augmentation de la fréquence des trains en Charente et Charente Maritime

### 3. Discussions

#### 3.1. Représentativité des contributeurs

Les contributions du Grand Débat National ne sont pas représentatives de la population française. En effet, les résultats que nous présentons dans cet article reflètent l'opinion des contributeurs, mais ne peuvent en aucun cas être considérés comme un reflet fidèle de l'opinion de la population. La méthode d'enquête de ce débat le rend plus proche de la production participative (*crowdsourcing*) ou du réseau social, que du sondage. La différence entre les statistiques classiques et l'analyse des réseaux sociaux (et plus généralement l'analyse de contenu généré par les utilisateurs) est bien connue, mais son importance a néanmoins été reconnue même en présence de différents types de biais (Morstatter, Liu, 2017). Ainsi, les caractéristiques sociologiques des contributeurs (âge, sexe, catégorie socio-professionnelle, ...) n'ont pas été demandées et sont absentes dans le jeu de données. Il n'est donc pas possible de s'assurer que les contributeurs constituent un échantillon représentatif de la population. Seul le code postal, renseigné par le contributeur, permet d'avoir une idée de la répartition spatiale des contributeurs. La participation n'a pas été homogène sur le territoire national (Bennani *et al.*, 2019). On constate par exemple une sous représentation des grands pôles urbains (58,4 % de la population contre seulement 27 % des contributeurs) et une sur représentation du périurbain (24,6 % contre 47,6 %) et des autres communes (17 % contre 25,5 %). De plus, il

faut noter que les cartes qui ont été présentées ne montrent pas le nombre de contributions dans lesquelles les motifs ont été détectés.

### *3.2. Couverture des motifs détectés dans les contributions*

Plusieurs limitations importantes du travail proposé dans cet article se rapporte à la couverture de la méthode de détection de motifs. Premièrement, notre méthode s'appuie sur un ensemble de lexiques non-exhaustifs. Deuxièmement, l'utilisation d'une fenêtre glissante n'est pas suffisante pour détecter des motifs plus complexes. Troisièmement, nous ne détectons pas l'emploi de la négation. Au delà de l'aspect de la méthode, le modèle de la "proposition" s'appuyant sur (Liu, Zhang, 2012) et les motifs choisis ne couvre pas l'ensemble des motifs disponibles. Par conséquent, un travail d'annotation et d'exploration des données plus profond sont nécessaires à l'amélioration de ce travail. Des travaux sont menées pour explorer deux nouvelles méthodes d'extractions. Les premiers s'appuient sur l'utilisation de la grammaire de dépendance (ou *dependency parsing*) (Kübler *et al.*, 2009) dans la formulation de règles d'extraction. Les seconds se concentrent sur la construction d'un corpus annoté pour la création d'un modèle de classification.

## **4. Conclusion**

L'analyse des contributions en ligne au *Grand Débat National* et la combinaison de cette analyse avec des données publiques, permettent de construire des cartes pour identifier des zones dans lesquelles une demande d'ajout, de développement d'équipements et d'infrastructures est proposé. Ces contributions recueillies à l'échelle nationale sont porteuses d'informations exploitables pour les décideurs locaux. Si la précision spatiale de la localisation des contributeurs, les échelles spatiales les plus intéressantes nous semblent être l'aire urbaine et le département. Les résultats présentés dans cet article sont néanmoins à considérer avec prudence. En effet, il n'est pas possible de considérer les contributeurs au Grand Débat National comme un échantillon représentatif de la population. De plus, l'extraction de motifs que nous avons proposée est trop restrictive pour que nous puissions affirmer avoir détecté toutes les contributions pertinentes.

Les futurs travaux porteront sur l'amélioration de l'extraction d'opinions. Plus particulièrement, par la mise en place d'une campagne d'annotation, puis par la construction d'une ressource lexicale (ontologie) plus complète sur la thématique du transport. L'objectif est de produire un corpus annoté permettant d'extraire d'identifier de nouveaux motifs et/ou de produire des modèles automatiques en s'appuyant sur des méthodes d'apprentissage automatique.

## Bibliographie

- Altman N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, vol. 46, n° 3, p. 175–185.
- Bellet A., Denis P., Gilleron R., Keller M., Vauquier N. (2020, juin). *Pour plus de transparence dans l'analyse automatique des consultations ouvertes : leçons de la synthèse du Grand Débat National*. Consulté sur <https://hal.inria.fr/hal-02860659> (working paper or preprint)
- Bennani H., Gandré P., Monnery B. (2019). *Les déterminants locaux de la participation numérique au grand débat national : une analyse économétrique*. Consulté sur [https://economix.fr/pdf/dt/2019/WP\\_EcoX\\_2019-7.pdf?1.0](https://economix.fr/pdf/dt/2019/WP_EcoX_2019-7.pdf?1.0)
- Boyer P. C., Delemotte T., Gauthier G., Rollet V., Schmutz B. (2020). Les déterminants de la mobilisation des Gilets jaunes. *Revue économique*, vol. 71, n° 1, p. 109–138.
- Breiman L. (2001). Random forests. *Machine learning*, vol. 45, n° 1, p. 5–32.
- Devlin J., Chang M.-W., Lee K., Toutanova K. (2019). *Bert: Pre-training of deep bidirectional transformers for language understanding*.
- Gerard Salton M. M. (1983). *Introduction to modern information retrieval*. McGraw-Hill.
- Kübler S., McDonald R., Nivre J. (2009). Dependency parsing. *Synthesis lectures on human language technologies*, vol. 1, n° 1, p. 1–127.
- Liu B., Zhang L. (2012). A survey of opinion mining and sentiment analysis. In *Mining text data*, p. 415–463. Springer.
- Loh W.-Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, n° 1, p. 14–23.
- Manning C. D., Schütze H., Raghavan P. (2008). *Introduction to information retrieval*. Cambridge university press.
- Messer M. A., Utz S., Moreau D. (2019). *Analyse exploratoire des contributions au grand débat national, volet mobilité*. techreport. Forum Vies Mobiles, Paris (France). Consulté sur [https://fr.forumviesmobiles.org/sites/default/files/editor/granddebat\\_rapport-final.pdf](https://fr.forumviesmobiles.org/sites/default/files/editor/granddebat_rapport-final.pdf)
- Morstatter F., Liu H. (2017, 1<sup>er</sup> 6). Discovering, assessing, and mitigating data bias in social media. *Online Social Networks and Media*, vol. 1, p. 1–13.
- Opinion Way. (2019). *Traitement des données issues du grand débat national*. Rapport technique. Auteur. Consulté sur <https://granddebat.fr/pages/syntheses-du-grand-debat>
- Ploux S., Genay M., Ploux-Chillès L. (2020, janvier). *Les mots du Grand débat national : quelques résultats et réflexions préliminaires*. Consulté sur <https://hal.archives-ouvertes.fr/hal-02441477> (working paper or preprint)
- Vapnik V. (2013). *The nature of statistical learning theory*. Springer science & business media.