Coronavirus - Étude de l'intensité médiatique

Nicolas Hervé
Institut National de l'Audiovisuel - Service de la Recherche
nherve@ina.fr

31 mars 2020 - v1.1

Historique

Version	Date	Auteur(s)	Modifications
1.0	24 mars 2020	NH	création, corpus jusqu'au 22 mars 2020 inclus
1.1	31 mars 2020	NH	ajout France 24 et radios (pages 4, 15, 19 et 20), étude rapide chloroquine / Didier Raoult (page 23), corpus jusqu'au 29 mars 2020 inclus

1 But de l'étude

Le but de cette étude est d'observer l'ampleur de la médiatisation du coronavirus sur différents supports en France (TV, radio, agence de presse, presse et Twitter) et de la mettre en relation avec les événements clés de la chronologie de cette épidémie ainsi qu'avec un certain nombre de données extérieures (nombre de cas, cours de bourse). En dehors d'un éclairage sur le système médiatique français, la compréhension de la circulation des informations en cas d'épidémie fait également partie des modélisations de propagation de la maladie [14] [16].

Nous commençons par présenter nos jeux de données et les approches quantitatives mises en œuvre aux chapitre 2 et 3. Les résultats bruts sont présentés au chapitre 4 en page 9. L'analyse de ces résultats mis en regard des données externes ainsi que de la chronologie des évènements est au chapitre 5, à partir de la page 20. Des graphiques complémentaires sont en annexe en fin de document.

Cette étude sera mise à jour aussi régulièrement que possible avec les nouvelles données ainsi qu'avec de nouveaux axes d'analyse. La dernière version ainsi que l'historique de ce document sont disponibles en ligne [1].

Les articles basés sur cette étude et publiés par La Revue des Médias sont les suivants :

— 24 mars, Information à la télé et coronavirus : l'INA a mesuré le temps d'antenne historique consacré au Covid-19 [Voir en ligne]
— 31 mars, Comment Didier Raoult et la chloroquine ont surgi dans le traitement médiatique du coronavirus [Voir en ligne]

Des travaux proches ont déjà été réalisés, nous listons ici ceux dont nous avons connaissance :

— 8 et 15 mars, Des posts sur Twitter par @Damien_Liccia [ici] et [ici]
— 16 mars, COVID-19 : The First Public Coronavirus Twitter Dataset [1]
— 20 mars, Covid-19 : histoire d'une médiatisation, par Le Temps et l'EPFL en Suisse [12]
— 20 mars, Études des commentaires Facebook par France Inter [6]
— 21 mars, Étude sur les article de presse, par le JDD [10]
— 26 mars, Twitter concentre 75% des conversations en ligne sur le coronavirus, par CBNews [9]
— 27 mars, Didier Raoult est devenu une star du web, par France Inter [3]

1. http://www.herve.name/pmwiki.php/Main/Etude-Coronavirus

2 Données

La particularité de notre approche est que nous captons en permanence les contenus médiatiques et Twitter en France. Cela nous offre un avantage précieux pour des études de ce type. En effet nous évitons ainsi deux des principaux écueils de la constitution de corpus d'actualité. D'une part, nous disposons de l'antériorité des documents, il n'est donc pas nécessaire de chercher à reconstituer l'historique d'un événement médiatique au moment où la décision est prise de l'étudier. D'autre part, nous essayons de capter de façon exhaustive les contenus que nous ciblons. Nous évitons ainsi les biais potentiels (souvent via la création de requêtes adéquates) liés à la constitution même des corpus qui sont, trop souvent, ignorés dans les analyses qui peuvent en être faites. Les corpus assemblés pour cette étude débutent au 1er décembre 2019. Même si aucune médiatisation n'est évidemment disponible pour le mois de décembre, nous utiliserons cette période comme jeu de contrôle pour les méthodes que nous appliquerons pour analyser le reste du corpus. Tout manque ou erreur dans ces données ou leur traitement sont de la seule responsabilité de l'auteur.

2.1 Données AFP

Les dépêches AFP ainsi que les articles de presse sont captés par la plateforme OTMedia développée au service de la Recherche à l'Ina [4]. À ce jour, seules les dépêches AFP sont intégrées dans l'étude. Nous ne conservons que les dépêches en français qui concernent un évènement en particulier. Aussi, toutes les dépêches concernant des prévisions de publication ou des agendas ne sont pas prises en compte dans cette étude. Ces dépêches spécifiques sont identifiées grâce à leurs titres qui commencent systématiquement par certains mots. Nous les présentons dans le tableau 2. Le tableau 3 synthétise les informations sur ce corpus de dépêches AFP, une fois supprimées celles que nous avons choisies d'ignorer.

agenda	l essentiel de l actualite
previsions	le monde en bref
a la une	en attendant demain
a noter pour aujourd hui	apres demain

Table 2 – Débuts des titres des dépêches AFP ignorées

Dates	Nb. jours	Nb. total docs.	Nb. moyen docs. par jour	Nb. mots	Nb. moyen mots par doc.
2019-12-01 → 2020-03-29	**119**	**118 774**	**998.45**	**45 481 555**	**382.93**

Table 3 – Informations sur le corpus de dépêches AFP

La figure 1 montre la répartition temporelle de ces dépêches. On observe clairement que moins de dépêches sont publiées lors des week-end, représentés sur un fond légèrement bleuté pour bien distinguer ces périodes. On observe également la baisse d'activité liée aux fêtes de fin d'année 2019.

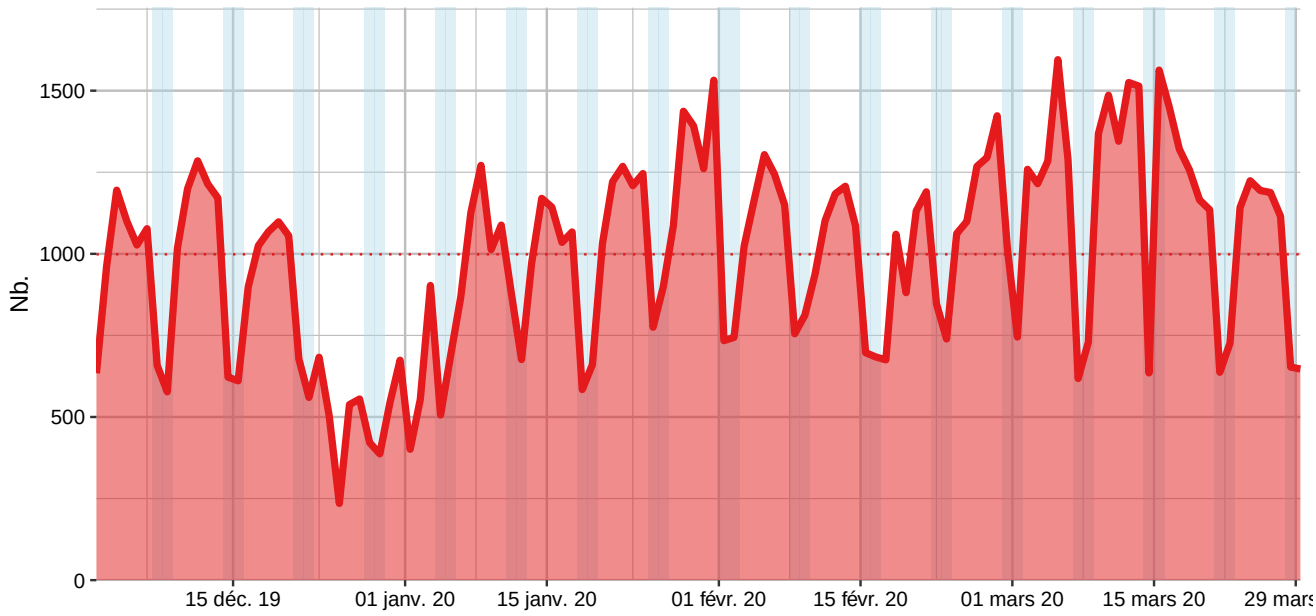


FIGURE 1 – Répartition temporelle des dépêches AFP ainsi que nombre moyen par jour. Les week-end sont représentés en bleu clair.

2.2 Données audiovisuelles

Pour la télévision, nous avons constitués deux corpus. Le premier contient les diffusions chaque jour des 5 chaînes d'information en continu entre 6h00 et minuit (BFMTV, CNews, LCI, franceinfo:et France 24). Le second corpus regroupe les plages d'information du soir sur les principales autres chaînes. Les créneaux concernés ont été déterminés avec les documentalistes de l'Ina [2]. Tous les jours de la semaine sont analysés, sauf le dimanche pour France 5. Pour la radio, seule France Info est traitée de la même façon que les chaînes TV d'information en continu (plage intégrale entre 6h00 et minuit). Pour les autres stations, c'est cette fois-ci la tranche matinale qui a été privilégiée. Le point important pour notre étude est que ces créneaux sont fixes. Nous ne tenons donc pas compte des éventuelles émissions spéciales ou de journaux TV particulièrement longs. De la même manière, les coupures de publicité ou les avant-programmes sont parfois dans le corpus. Le tableau 4 synthétise les informations sur les créneaux TV et radio que nous avons inclus dans notre corpus [3].

Pour traiter ces données, nous commençons par extraire le texte de tout ce qui est prononcé à l'antenne. Cette transcription des flux audio est assurée à l'aide du logiciel développé par le laboratoire du LIUM de l'Université du Mans [15] déployé sur nos serveurs. Comme tous les logiciels de ce type, la transcription est basée sur un dictionnaire de termes et sur un modèle de la langue française. Les résultats obtenus sont de bonne qualité mais ne peuvent pas être parfaits. Les erreurs peuvent être dues à deux raisons principales :

- les conditions acoustiques sont mauvaises : la transcription d'une émission en plateau télé sera de meilleure qualité qu'un micro-trottoir ou qu'une conversation téléphonique
- les termes employées ne sont pas connus du logiciel : c'est principalement le cas avec certains noms propres ou avec de nouveau termes. Ainsi le logiciel du LIUM, avec le modèle dont nous disposons qui date de quelques mois, connaît bien le mot *coronavirus* mais pas le terme *Covid*. Ce dernier peut donc être transcrit sous différentes formes qui sont phonétiquement proches. En voici quelques exemples trouvés dans les résultats : co vide, koweït, code vide, comite, covic, lukovic, aucun vide, coville, ...

2. Pour TF1 et France 2, le JT du soir était initialement transcrit entre 19h55 et 20h40, la durée a finalement été étendue à la plage horaire 19h55 - 21h05 suite à l'annonce de France 2 de rallonger son JT à partir du 15 mars.

3. Les données sont manquantes pour l'instant pour le créneau 14h00→16h00 le 9 décembre 2019 pour la radio France Info

	Chaîne / Station	Émission	Jours	Durée transcrite	Nb. mots	Nb. moyen mots par heure
Télévision	BFM TV		6h00→0h00	**2 158** h	**24 062 575**	**11 150**
	franceinfo:		6h00→0h00	**2 158** h	**23 189 026**	**10 746**
	CNews		6h00→0h00	**2 158** h	**24 050 970**	**11 145**
	LCI		6h00→0h00	**2 158** h	**25 310 084**	**11 728**
	France 24		6h00→0h00	**2 158** h	**21 512 277**	**9 969**
	TF1[2]	JT de 20h00	19h55→21h05	**142** h	**1 400 496**	**9 863**
	France 2[2]	JT de 20h00	19h55→21h05	**142** h	**1 386 342**	**9 763**
	France 3	JT de 19/20	19h25→20h05	**80** h	**753 151**	**9 414**
	France 5	C dans l'air	17h40→18h50	**119** h	**1 480 456**	**12 441**
	M6	JT du soir	19h40→20h15	**70** h	**701 847**	**10 026**
	Total TV			**11 343** h	**123 847 224**	**10 918**
Radio	France Info[3]		6h00→0h00	**2 156** h	**24 533 724**	**11 379**
	France Inter	Matinale	4h55→9h05	**500** h	**5 249 149**	**10 498**
	Europe 1	Matinale	6h55→09h05	**260** h	**3 247 133**	**12 489**
	RFI	Matinale RFI Afrique	7h55→8h20 20h25→20h45	**90** h	**1 019 612**	**11 329**
	RMC	Matinale	19h25→20h05	**380** h	**4 721 100**	**12 424**
	RTL	Matinale	19h25→20h05	**320** h	**3 941 106**	**12 316**
	Total Radio			**3 706** h	**42 711 824**	**11 525**
Total corpus audiovisuel				**15 049** h	**166 559 048**	**11 068**

Table 4 – Corpus des créneaux TV et radio

2.3 Données Twitter

Les tweets sont captés à l'aide du logiciel mis en place par Béatrice Mazoyer dans le cadre de sa thèse [8]. Le principe général est de permettre une captation continue de tweets en français en se basant sur des requêtes de mots neutres les plus couramment utilisés (*stop words*). L'avantage de cette approche est qu'elle permet de capter, via l'API fournie par Twitter, un volume de tweets suffisants pour réaliser des études statistiques sans avoir à définir au préalable des termes de recherche et en garantissant une distribution des tweets équivalente à celle fournie par l'API sample [4]. Nos estimations, selon différentes approches, nous conduisent à penser que nous captons environ 60% des tweets en français émis sur la plateforme de micro-blogging. Sur la période nous captons environ 5.5 millions de tweets par jour. Il s'agit bien de tweets en français et non uniquement de tweets émis depuis la France. Nous n'utilisons pas les informations de géolocalisation des comptes Twitter ou des tweets, trop parcimonieuses et imprécises. Les détails sont donnés dans le tableau 5. Il faut toutefois noter que cette estimation est valable globalement pour une longue période. Ponctuellement, notamment en cas de période de forte publication, le pourcentage de tweets captés peut être plus faible. L'API de Twitter plafonne en effet le volume de tweets qui peuvent être récupérés à un instant donné. C'est pourquoi les volumes de tweets sont donnés dans cette étude à titre indicatif. On constate par exemple sur le graphique 2 une légère baisse d'activité à Noël et un net regain après le 1er de l'an. Un pic est également observé depuis mi-mars. Pour ces deux pics, nous ne savons pas s'ils ont été écrêtés par Twitter ou s'ils correspondent toujours à environ 60% des tweets émis en français. On peut toutefois raisonnablement penser qu'il y a un regain d'activité sur Twitter à ces deux périodes. Il conviendra donc de se concentrer plutôt sur les ratios de tweets qui sont plus pertinents et non biaisés.

4. API fournissant 1% du contenu Twitter, choisi selon des critères qui ne sont pas connus mais que l'on sait ne pas être biaisés et ainsi fournir un échantillon représentatif

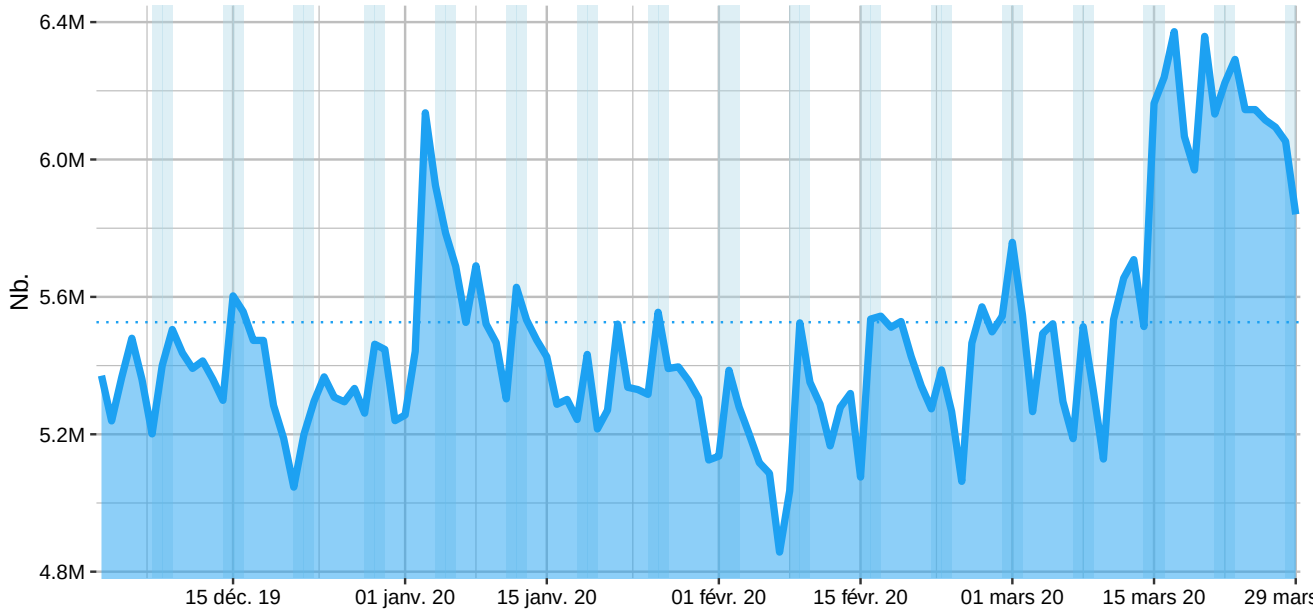


FIGURE 2 – Répartition temporelle des tweets captés ainsi que nombre moyen par jour.

Dans le tableau 5, le nombre de mots des tweets est calculé sur le texte obtenu de la façon suivante [5]. Il s'agit du pré-traitement spécifique aux tweets pour notre étude.

- le texte du tweet est récupéré
- s'il s'agit d'une réponse à un tweet (*reply*), le texte du tweet initial est également ajouté
- s'il s'agit d'une citation d'un tweet (*quote*), le texte du tweet initial est également ajouté
- sur le texte ainsi constitué, on supprime ensuite quelques mots spécifiques :
  - les mentions à des comptes Twitter (mots commençant par un @)
  - les URLs (mots commençant par http)
  - l'indication de retweet est supprimée si elle est présente (mot *RT*)
  - les emojis

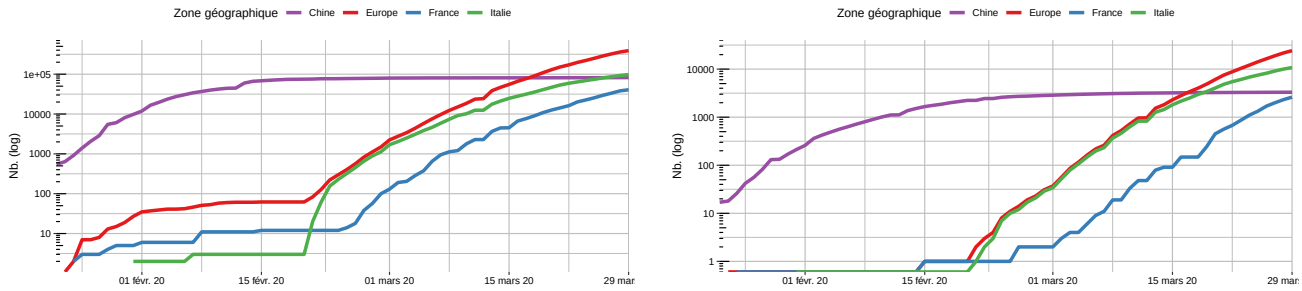
Dates	Nb. jours	Nb. total tweets	Nb. moyen tweets / jour	Nb. mots	Nb. moyen mots / tweet
2019-12-01 → 2020-03-29	**119**	**657 630 628**	**5 526 308**	**20 746 724 318**	**31.55**[5]

TABLE 5 – Informations sur le corpus de tweets

2.4 Données externes

Les données épidémiologiques sur le coronavirus sont agrégées et mises à disposition tous les jours par le Center for Systems Science and Engineering de l'Université John Hopkins [2]. Les données ne sont disponibles qu'à partir du 22 janvier 2020. Elle concernent le nombre de cas confirmés ainsi que le nombre de décès officiellement attribués au coronavirus. Les données pour la Chine concernent les séries *Mainland China* et *China* dans les données du CSSE. Pour la France, nous n'avons pris que les données de la métropole. Pour l'Europe, nous avons choisi de représenter la zone géographique plutôt que l'instance politique, qui s'étend donc jusqu'à la Russie incluse. On représente les principaux chiffres dans la figure 3. Ce graphique utilise une échelle logarithmique, permettant de bien visualiser la propagation exponentielle du virus en Europe.

5. Le nombre moyen de mots par tweet peut sembler particulièrement élevé pour celles et ceux qui utilisent régulièrement Twitter. Nous rappelons que nous prenons en compte le texte du tweet mais également celui des *reply* et *quote*.



(a) Cas (b) Décès

FIGURE 3 – Données épidémiologiques agrégées par JH CSSE

Les cours de bourse sont disponibles en ligne [6]. Nous avons choisi les principaux indices pour la France (CAC40), l'Italie (FTSE MIB), la Chine (Shanghai Composite) et les États-Unis (Dow Jones) ainsi que le cours du baril de pétrole brut(Brent). La figure 4 présente l'évolution de ces cours en prenant pour base 100 la valeur de clôture au 31 décembre 2019, soit la dernière valeur disponible avant que la Chine ne contacte officiellement l'OMS pour prévenir de l'existence du nouveau coronavirus.

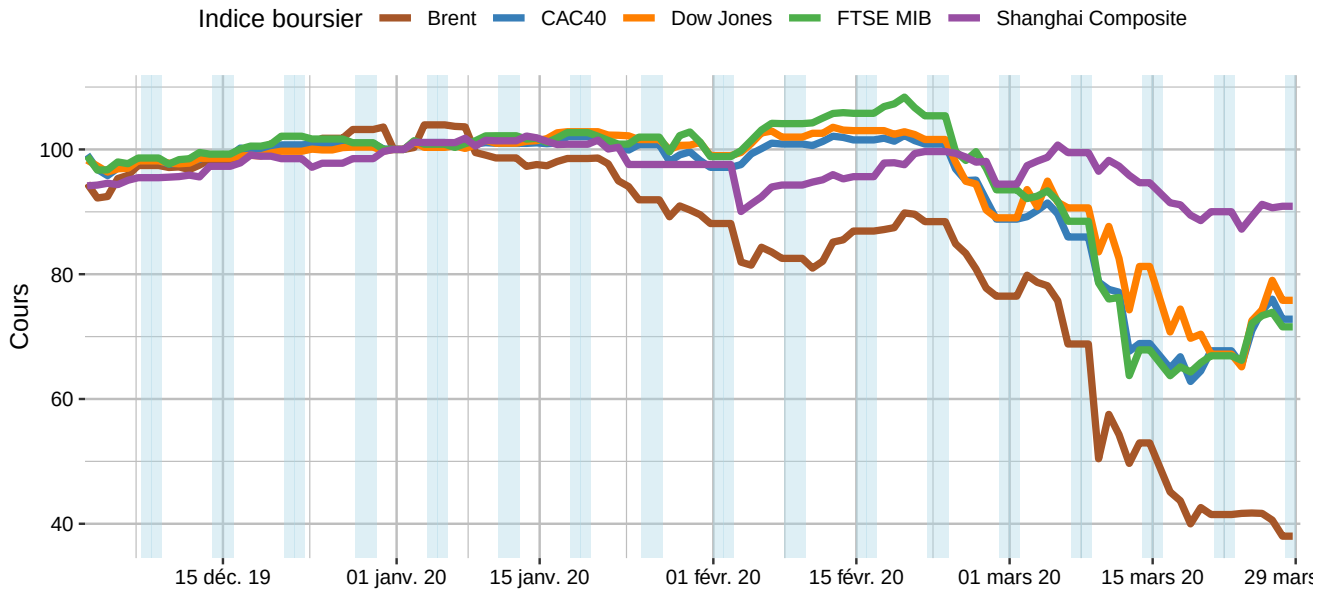


FIGURE 4 – Valeurs des cours de bourse, base 100 au 1er janvier 2020

3 Méthodologie

Nous mettons en place différents algorithme de quantification automatique se basant sur nos jeux de données pour extraire les informations pertinentes permettant une analyse de la médiatisation du coronavirus sur les différents supports en France.

3.1 Textométrie

Nous déterminons un ensemble de mots caractéristiques permettant de cerner certains aspects de l'épidémie et de sa médiatisation. La découverte de ces mots est réalisée à partir d'une étude de leur fréquence et de leur pertinence (utilisation de TF-IDF [13]) et il sont validés manuellement. Ces mots

6. https://www.boursorama.com/bourse/indices/internationaux

sont ensuite répartis en groupes selon leurs thématiques. Cette répartition, forcément subjective, est susceptible d'évoluer au fur et à mesure de l'actualité ou des raffinements successifs de cette étude.

Nous comptons ensuite toutes les occurrences d'apparition de ces mots dans les textes (dépêches d'agence, presse en ligne et tweets) ainsi que dans les transcriptions (flux TV). Tous les textes traités sont préalablement normalisés (minuscules, suppression des accents et des caractères non-alphanumériques). Cette approche textométrique simple permet déjà d'observer quelques phénomènes. Nous présentons dans le tableau 6 l'ensemble des mots utilisés. Il s'agit plus précisément des préfixes de mots. Ainsi par exemple le préfixe *confine* permet de compter tous les mots qui commencent par *confine*, par exemple *confine*, *confiner*, *confinés*, *confinement*, ... Deux groupes de vocabulaire sont liés aux zones géographiques chinoise et italienne qui ont été importante dans le démarrage de l'épidémie. Les quatre autres groupes concernent le virus proprement dit, les questions médicales, l'éducation ainsi que les mesures prises pour lutter contre l'épidémie. Nous définissons également le supra-groupe *coronavirus* qui englobe *virus*, *médecine* et *mesures*. Certains aspect de l'épidémie ne sont pas couverts par notre vocabulaire, notamment toutes les questions économiques (conséquences, mesures spécifiques, continuité d'activité, ...) ainsi que les élections municipales.

groupe	liste des préfixes
chine	chine, chinois, hubei, pekin, shanghai, wuhan
education	baccalaureat, cned, college , colleges , collegien, ecole, ecoli, educ, eduqu, eleves, enseign, etudi, lycee, maternell, periscolair, professeur, scolair, scolaris, universit
italie	italie, italien, lombardie, milan, peninsule, rome, venetie, venise
medecine	asymptomat, chercheur, chirurgi, cliniqu, depistag, diagnost, docteur, ehpad, hopita, hospital, immunise, immunitaire, infirmier, laboratoire, lit, malad, medecin, medica, oms, pathologie, patient, pharmac, pulmonaire, reanim, recherche, respirat, samu, sanitaire, sante, scientifique, soign, symptom, traitement, urgenc, vaccin
mesures	annul, barriere, chez vous, confine, coude, ferm, fermeture, gel, geste, hydroalcoolique, les mains, masque, quarantaine, rapatrie, report , reporte, savon, stade trois, teletravail, usage unique
virus	cas confirme, contagi, contamination, contamine, corona, covid, desinfect, dix neuf, epidemi, fievr, forme severe, formes severes, gripp, infect, pandemi, pneumonie, propag, respiratoir, sars, sras, transm, viral, virolog, virus
coronavirus	ce groupe spécial cumule tous les mots de *virus*, *mesures* et *medecine*

Table 6 – Vocabulaire - groupes de mots utilisés (préfixes)

3.2 Détermination du temps d'antenne

Nous avons déjà utilisé une approche permettant, sur la base d'un vocabulaire, d'estimer le temps d'antenne dédié au traitement d'un sujet. Elle a été mise en œuvre dans le cadre des études comparant la médiatisation de deux événements concomitants : Marche climat vs. Gilets jaunes [11] et Chirac vs. Lubrizol [7]. Nous détaillons ici plus précisément le fonctionnement de cet algorithme. Il s'agit globalement d'une estimation de densité d'apparition des mots du vocabulaire sur la transcription du flux audio. Intuitivement, plus on observe d'apparition de mots dans une courte période, plus la probabilité est élevée que cette période parle de la thématique en question. Un groupe virtuel *autre* est utilisé pour modéliser les autres thématiques de l'actualité qui ne seraient pas représentées dans notre vocabulaire. L'algorithme 1 est décrit de façon simplifiée en pseudo-code (pour une meilleure lisibilité, nous ne faisons notamment pas apparaître les cas pour lesquels les variables sont égales à 0 ni la gestion des erreurs). Les paramètres de cet algorithme (durée de la fenêtre, gaussienne, seuils) ont été déterminés et validés à la suite d'une comparaison avec un décompte humain des temps d'antenne lors de la première étude [11]. Ce sont les mêmes qui sont utilisés depuis. Seul le vocabulaire varie. Ces résultats automatiques comportent quelques biais connus et maîtrisés. La principale différence entre les deux précédentes études et celle-ci est la durée de l'événement observé. En effet, plus un événement dure dans le temps, plus son traitement médiatique va s'attarder sur ses différents aspects et donner lieu à des choix éditoriaux et des angles de traitement de l'actualité. Ainsi, lors de la seconde étude [7] nous avons estimé que les chiffres

Données : une transcription de durée d secondes, un vocabulaire composé de groupes de mots
Résultat : estimation des segments liés à chaque groupe de mots du vocabulaire

`/* Mixture de gaussiennes */`

$g \leftarrow \mathcal{N}(\mu = 0, \sigma^2 = 20s)$
pour *chaque groupe du vocabulaire* **faire**
  pour $s \in [0, d[$ **faire**
    $gmm[groupe][s] \leftarrow 0$

`/* Parcours de la transcription avec une fenêtre glissante */`

$debutFenetre \leftarrow 0s$
tant que $debutFenetre < d$ **faire**
  $finFenetre \leftarrow debutFenetre + 60s$
  `/* Comptage des mots */`
  $nbMotsTotal \leftarrow$ nombre total de mots total dans la fenêtre
  $nbMotsTotalNrm \leftarrow log(nbMotsTotal)$
  $nbMotsTrouve \leftarrow 0$
  pour *chaque groupe du vocabulaire* **faire**
    $nbMotsGroupe \leftarrow$ nombre de mots du groupe dans la fenêtre
    $nbMotsTrouve \leftarrow nbMotsTrouve + nbMotsGroupe$
  `/* Softmax pour les groupes en concurrence sur le même segment */`
  $expMots \leftarrow exp(nbMotsTrouve)$
  $expAutre \leftarrow exp(nbMotsTotalNrm - nbMotsTrouve)$
  $smMots \leftarrow expMots/(expMots + expAutre)$
  pour *chaque groupe du vocabulaire* **faire**
    $nbMotsGroupe \leftarrow$ nombre de mots du groupe dans la fenêtre
    $softmax[groupe] \leftarrow smMots * nbMotsGroupe/nbMotsTrouve$
  $softmax[autre] \leftarrow expAutre/(expMots + expAutre)$
  `/* Ajout à la mixture de gaussiennes */`
  pour *chaque groupe du vocabulaire* **faire**
    $v \leftarrow softmax[groupe] * g$
    pour $s \in [debutFenetre, finFenetre]$ **faire**
      $gmm[groupe] \leftarrow gmm[groupe] + gaussienneCentree(v, s)$
  $debutFenetre \leftarrow debutFenetre + 30s$

`/* Normalisation et seuillage par hystéresis */`

pour *chaque groupe du vocabulaire* **faire**
  $segmentation[groupe] \leftarrow hyteresis(gmm[groupe]/max(gmm), 0.3, 0.5)$

Algorithme 1 : Segmentation du temps d'antenne

concernant le temps d'antenne consacré au décès de Jacques Chirac étaient légèrement sous-estimés de l'ordre 5 %. Ceci était dû aux passages pendant lesquels des invités en plateau retraçaient la vie de l'ancien président. On pouvait assister à de longs monologues sans que son nom ne soit cité, le téléspectateur étant supposé connaître le contexte de l'intervention. Le choix du vocabulaire autour du coronavirus a donc été fait en tenant compte de cette contrainte et pourra d'ailleurs évoluer dans les futures versions de ce document. Nous avons cherché à distinguer les mots de vocabulaire liés spécifiquement à ce virus, au domaine médical, au mesures prises par les autorités et enfin aux conséquences. Tout ces mots peuvent évidemment apparaître dans l'actualité en dehors du contexte spécifique de la pandémie de coronavirus. C'est la raison pour laquelle nous conservons dans tous nos jeux de données le mois de décembre 2019. Il sert d'étalon pour mesurer les biais et avoir une bonne idée de la marge d'erreur de nos mesures pour la suite.

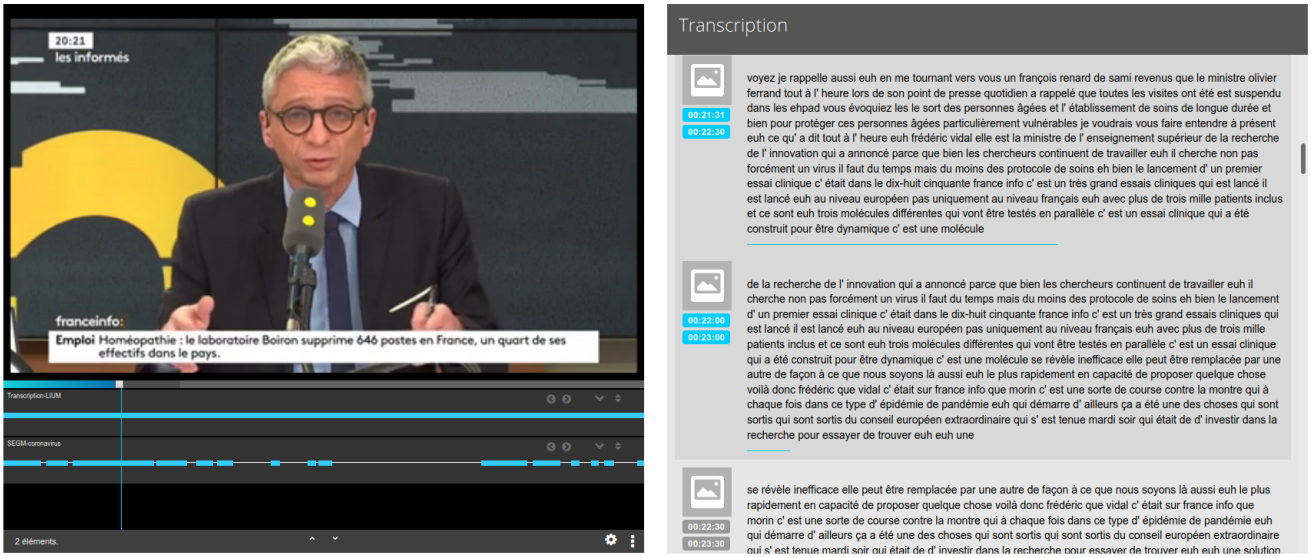


FIGURE 5 – Interface de visualisation des segmentations du temps d'antenne et des transcriptions, basée sur le player vidéo amalia.js [5]. Sur cet exemple, on a la tranche 20h00-22h00 du 11 mars 2020 de franceinfo:. Le long segment qui n'est pas lié au coronavirus est un débat en plateau sur les scandales politico-financiers.

4 Résultats bruts

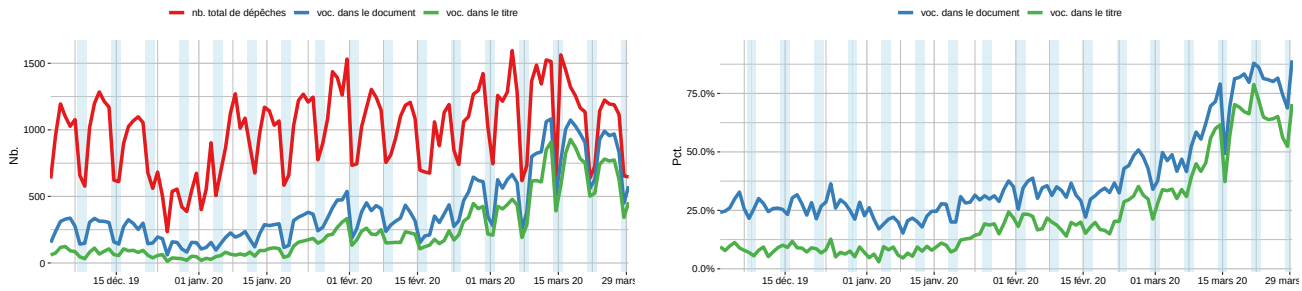
Nous présentons dans ce chapitre les résultats bruts issus des algorithmes de quantification automatique mis en œuvre sur les différents jeux de données.

4.1 Dépêches AFP

Les dépêches AFP ont un titre et un contenu. Pour déterminer si une dépêche aborde une des thématiques, nous considérons qu'il suffit qu'un seul mot apparaisse dans le titre. En revanche nous fixons la limite à 5 mots pour le contenu de la dépêche. Le nombre de dépêches qui respectent ces critères (pour l'ensemble des groupes de mots du vocabulaire) est présenté sur la figure 6.

Si on regarde maintenant la répartition des groupes de vocabulaire sur les dépêches dans lesquelles le titre est concerné, on observe que sur notre période de contrôle de décembre 2019 des dépêches concernent déjà nos thématiques. On indique dans le tableau 7 quelques exemples de dépêches du mois de décembre pour lesquelles le contenu est particulièrement en phase avec les thématiques que nous avons développées pour cette étude. Ces faux positifs sont une bonne indication des sujets d'actualités notamment liés à la question de la gestion des hôpitaux en France avec leur manque de moyen chronique. On observe de la même manière en annexe 29 que le groupe de mots *medecine* est particulièrement retrouvés dans les dépêches en décembre.

On observe clairement sur le graphique 7 trois principales étapes dans le traitement de l'épidémie par l'AFP, avec un nette augmentation de la part de dépêches qui y est consacrée à chaque fois. La



(a) Nombre de dépêches

(b) Pourcentage de dépêches

FIGURE 6 – Nombre et pourcentage de dépêches AFP pour lesquelles au moins un mot du vocabulaire est trouvé dans le titre ou 5 mots dans le corps du document.

date	Titre des dépêches AFP
2019-12-01	Rien n'arrête la bactérie qui tue les orangers de Floride (ou presque)
2019-12-03	Education : la Chine meilleure élève de l'étude Pisa, la France dans la moyenne
2019-12-04	La rougeole dévaste les Samoa
2019-12-04	Ultimes préparatifs avant un jeudi noir contre la réforme des retraites
2019-12-10	"On est au bord du gouffre" : à Nice, les internes entrent en grève illimitée
2019-12-12	Ebola en RDC : 20 nouveaux cas en trois jours, une nette reprise à la hausse
2019-12-13	L'AP-HP s'organise pour fonctionner pendant la grève, qui ajoute à la fatigue du personnel
2019-12-17	Blouses blanches et "pom pom girls" en études de médecine : l'hôpital dans la rue
2019-12-26	En 2019, le pas de géant de la thérapie génique

TABLE 7 – Faux positifs dans les dépêches AFP de décembre 2019

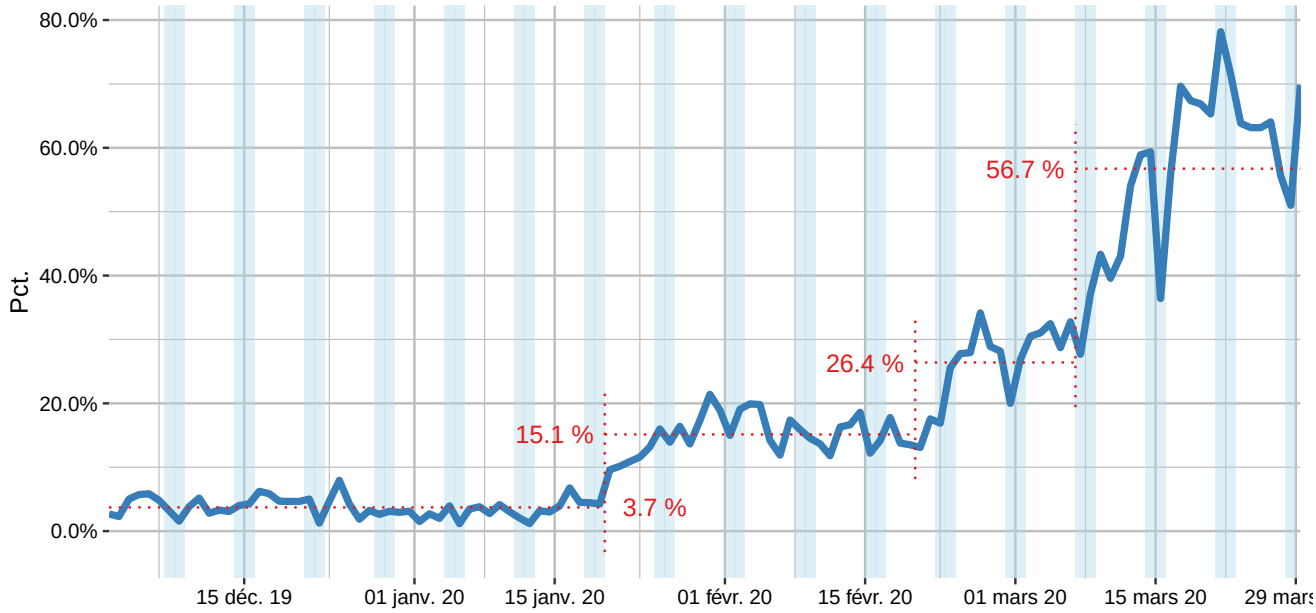


FIGURE 7 – Pourcentage de dépêches AFP pour lesquelles au moins un mot du groupe *coronavirus* est trouvé dans le titre et moyenne sur les différentes périodes.

dernière semaine de janvier 2020, c'est la Chine qui est principalement traitée. On a en moyenne 15% des dépêches évoquant le virus dans leur titre. Fin février, l'épidémie se développe fortement en Italie et on passe à un un ratio de l'ordre de plus de 25% de dépêches liées au coronavirus. Enfin, à partir de la deuxième semaine de mars, la France est plus durement touchée et on dépasse 50% de dépêches sur le sujet pour presque atteindre les 80%. Les détails pour chaque groupe de vocabulaire sont dans la figure

9. On remarque de plus la journée particulière du 15 mars, premier tour des élections municipales, avec un net recul pour ce seul jour de la thématique coronavirus dans les titres de dépêches. Enfin, dans la figure 29 en annexes nous présentons les valeurs brutes du nombre de mots des groupes de vocabulaires trouvés dans les documents AFP et les détails dans la figure 30.

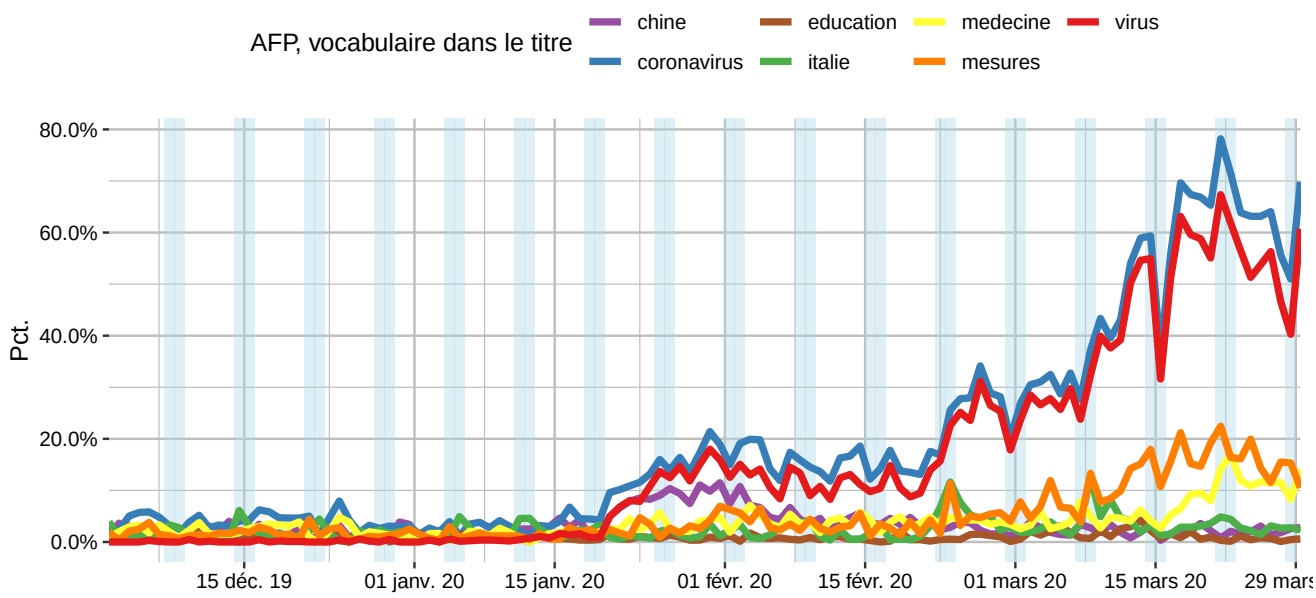


FIGURE 8 – Pourcentage de dépêches AFP pour lesquelles au moins un mot est trouvé dans le titre en fonction des groupes de vocabulaire.

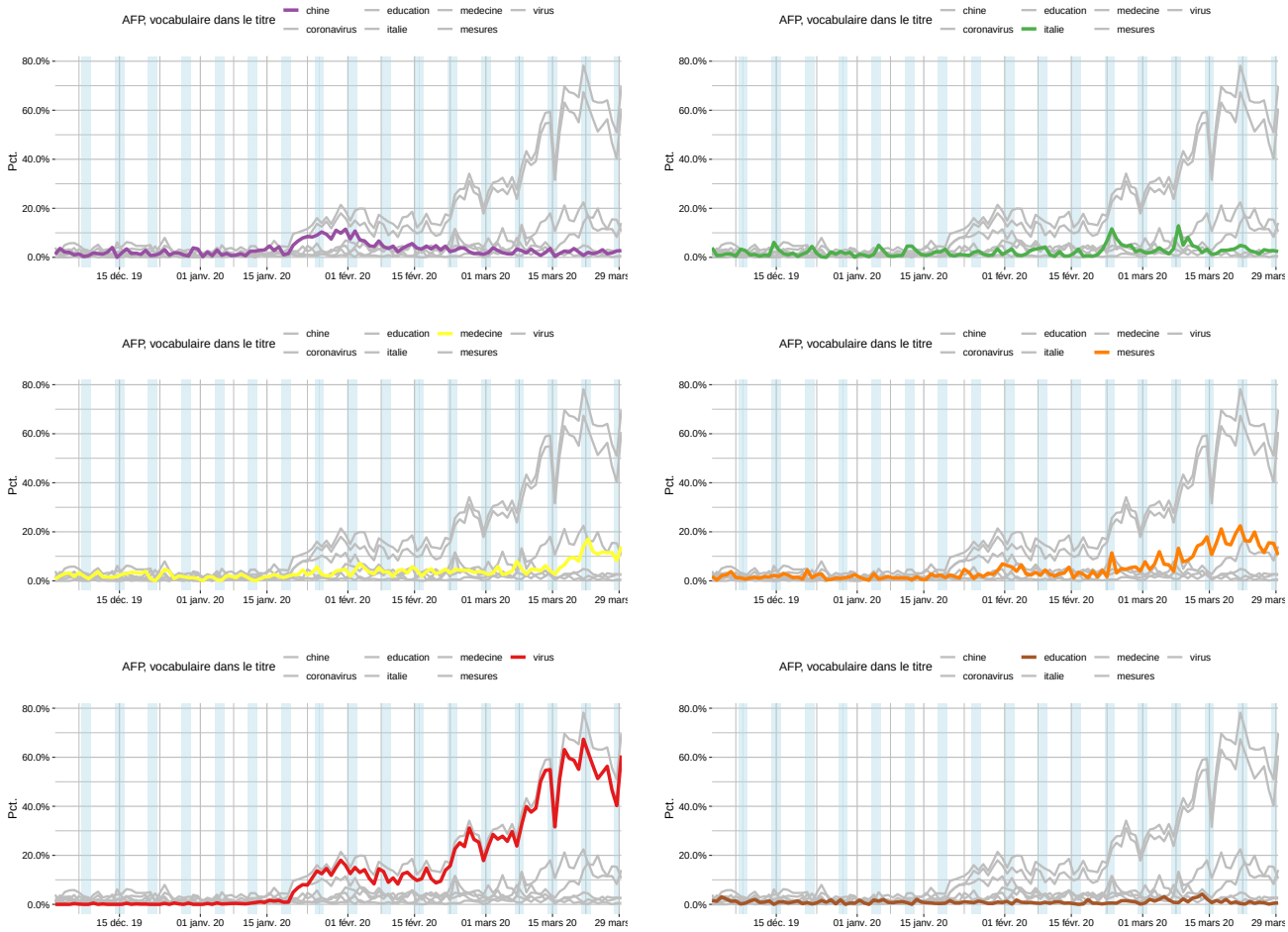
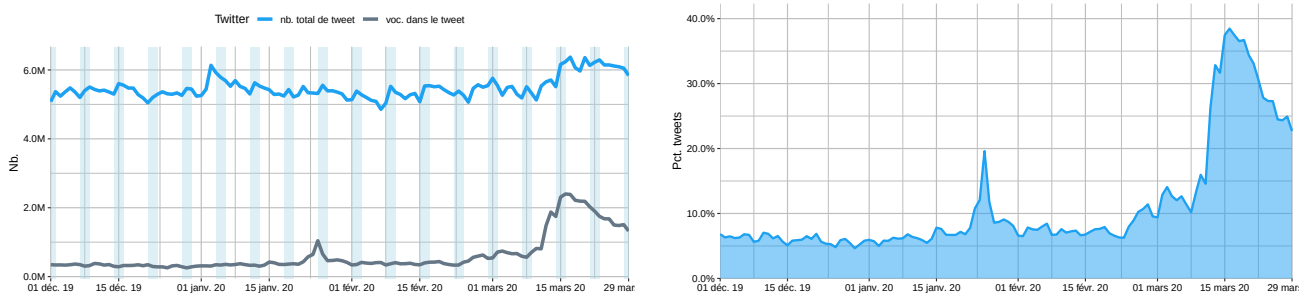


FIGURE 9 – Pourcentage de dépêches AFP pour lesquelles au moins un mot est trouvé dans le titre pour chaque groupe de vocabulaire

4.2 Twitter

Par rapport aux dépêches AFP, on observe sur Twitter un comportement beaucoup plus marqué. Un premier pic apparaît fin janvier avec la médiatisation de la situation en Chine et le premier cas français. Un léger palier est franchi début mars, mais c'est ensuite clairement à partir du 12 mars et de la première allocution télévisée d'Emmanuel Macron sur le sujet qu'un cap est franchi avec environ 35% des tweets en français qui évoquent le sujet.

Si on regarde plus précisément les groupes de vocabulaire (figure 12), on remarque depuis le 15 mars une baisse de l'utilisation des termes liés au virus et une nette augmentation de ceux liés aux mesures prises et, dans une moindre mesure, des termes médicaux. Les rumeurs de confinement circulent en effet pendant ce week-end d'élections et ce dernier est effectivement annoncé le lundi 16. Le contexte étant maintenant évident pour tout le monde, il n'est peut être plus nécessaire d'évoquer dans les tweets (où la place manque parfois) les termes liés au virus, et les récits du confinement deviennent plus présents. La situation dans les hôpitaux est également un sujet de discussion. La fermeture des établissements scolaires provoque un léger pic sur ce vocabulaire mais il retombe rapidement.



(a) Nombre de tweets

(b) Pourcentage de tweets

FIGURE 10 – Nombre et pourcentage de tweets en français captés pour lesquels au moins un mot du vocabulaire est trouvé.

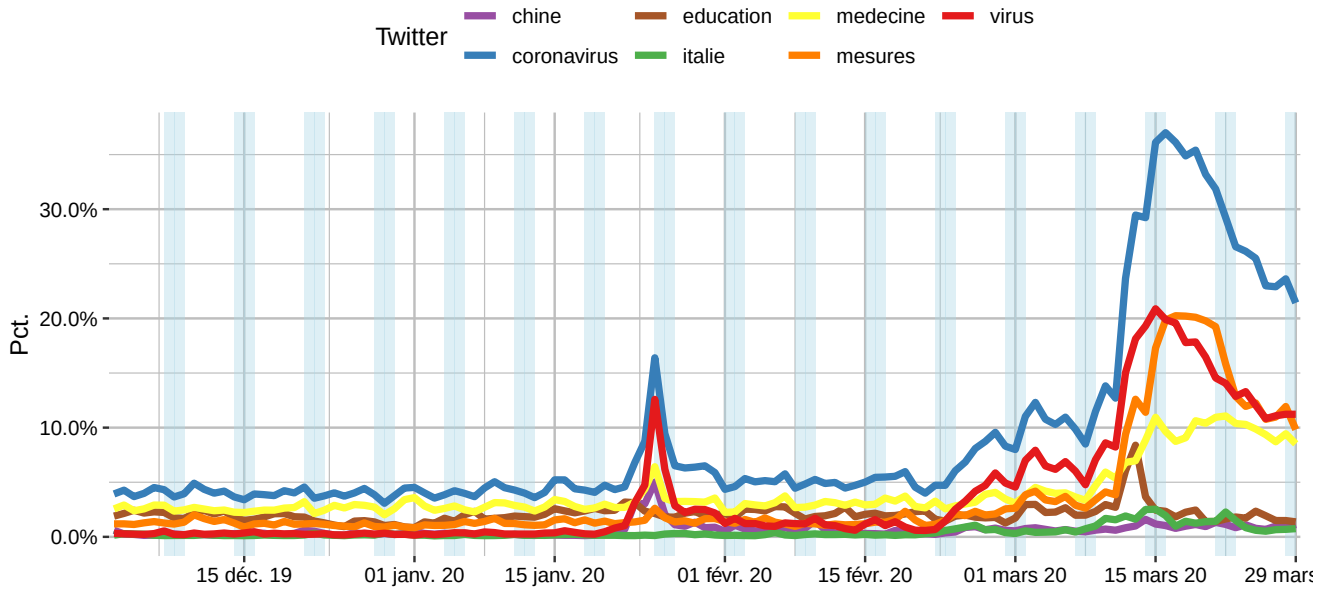


FIGURE 11 – Pourcentage de tweets en français captés pour lesquels au moins un mot du vocabulaire est trouvé en fonction des groupes de vocabulaire.

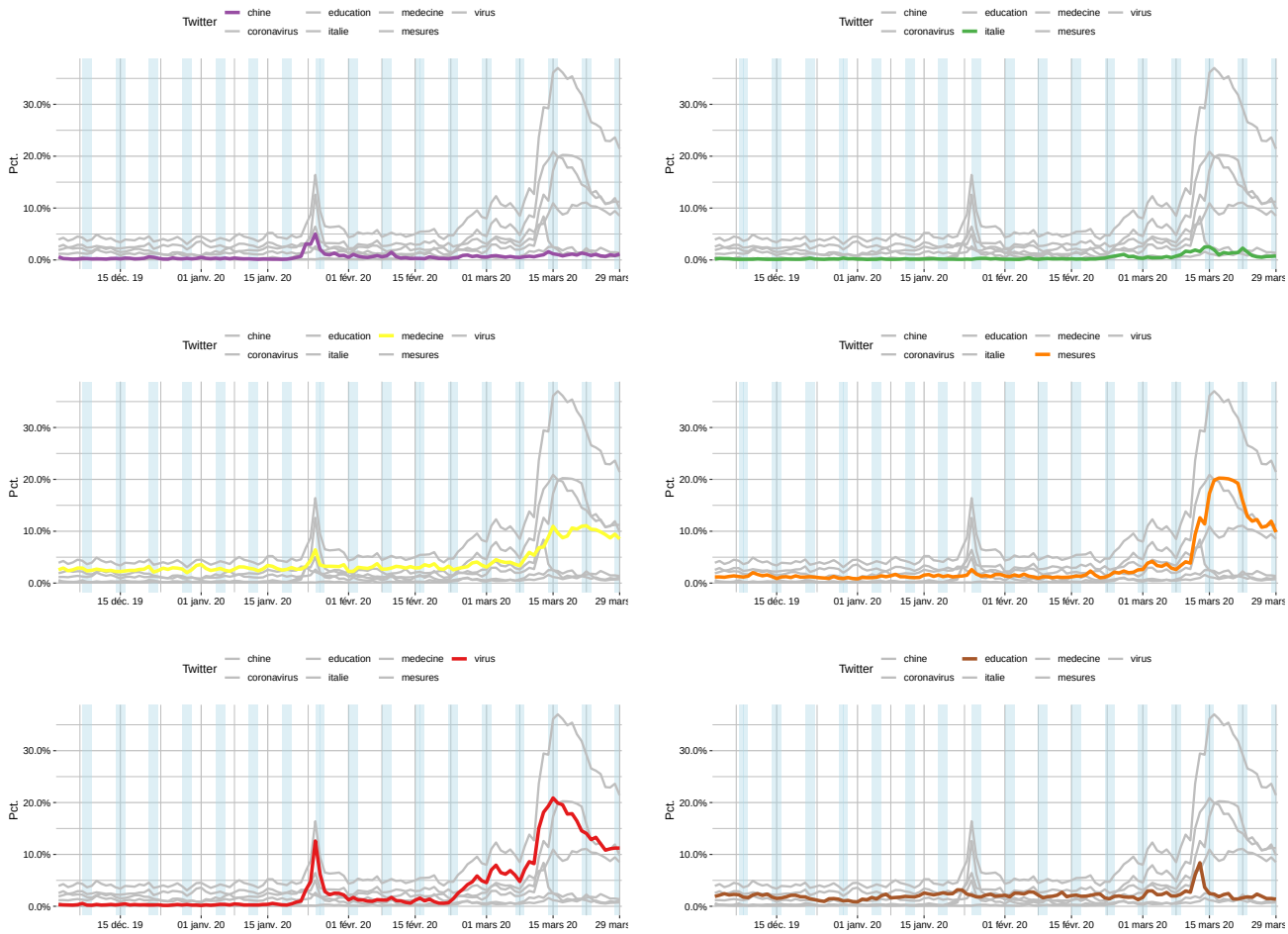


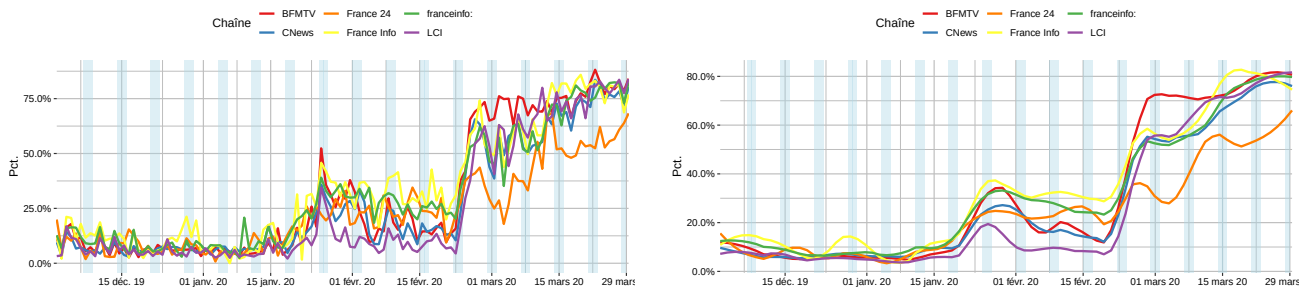
FIGURE 12 – Pourcentage des tweets en français captés pour lesquels au moins un mot du groupe de vocabulaire est trouvé

4.3 Information en continu

Les chaînes d'information en continu ont toute un profil similaire sur cette période. Pour mieux distinguer les tendances générales, on applique un *smoothing* local[7] sur la figure 13. Suite au premier pic de médiatisation fin janvier, franceinfo:consacre ainsi plus de deux fois plus de temps d'antenne au coronavirus que LCI. Le second pic de médiatisation à partir de mi-février est plus marqué en revanche sur BFMTV qui est rattrapé par les trois chaînes à partir de mi-mars. Seule France 24 semble sur cette dernière période consacrer un temps d'antenne moindre au coronavirus. Pour les chaînes privées, il ne faut pas oublier qu'il y a encore des plages de coupures de publicités. Outre des lacunes thématiques dans notre vocabulaire, ceci explique en partie le fait que nous n'ayons pas atteint les 100% de temps d'antenne sur ces chaînes.

Si on regarde le vocabulaire, on note une légère baisse des termes directement liés au *virus* depuis début mars. Comme pour Twitter, le contexte étant maintenant bien installé, il n'est plus la peine de répéter aussi souvent à l'antenne les mots du coronavirus. Les téléspectateurs ont bien intégré cela. En revanche, on observe sur la même période une montée des termes *medecine* et *mesures*.

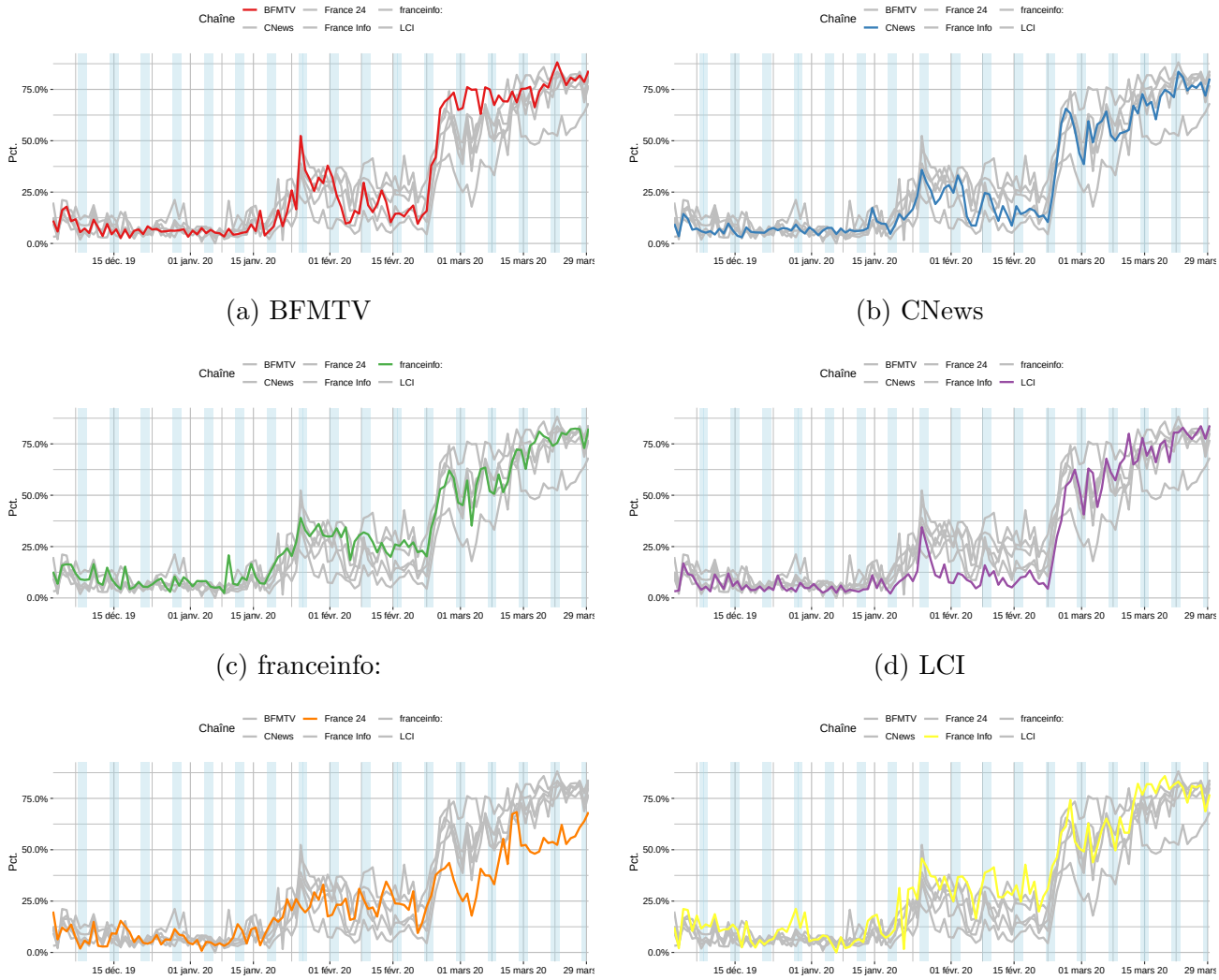
7. régression polynomiale, fonction loess(), span = 1.5 dans R



(a) Standard

(b) Avec *smoothing*

FIGURE 13 – Temps d'antenne estimé sur le coronavirus pour l'ensemble des canaux d'information en continu.



(a) BFMTV

(b) CNews

(c) franceinfo:

(d) LCI

(e) France 24

(f) France Info

FIGURE 14 – Temps d'antenne estimé sur le coronavirus pour chaque canal d'information en continu.

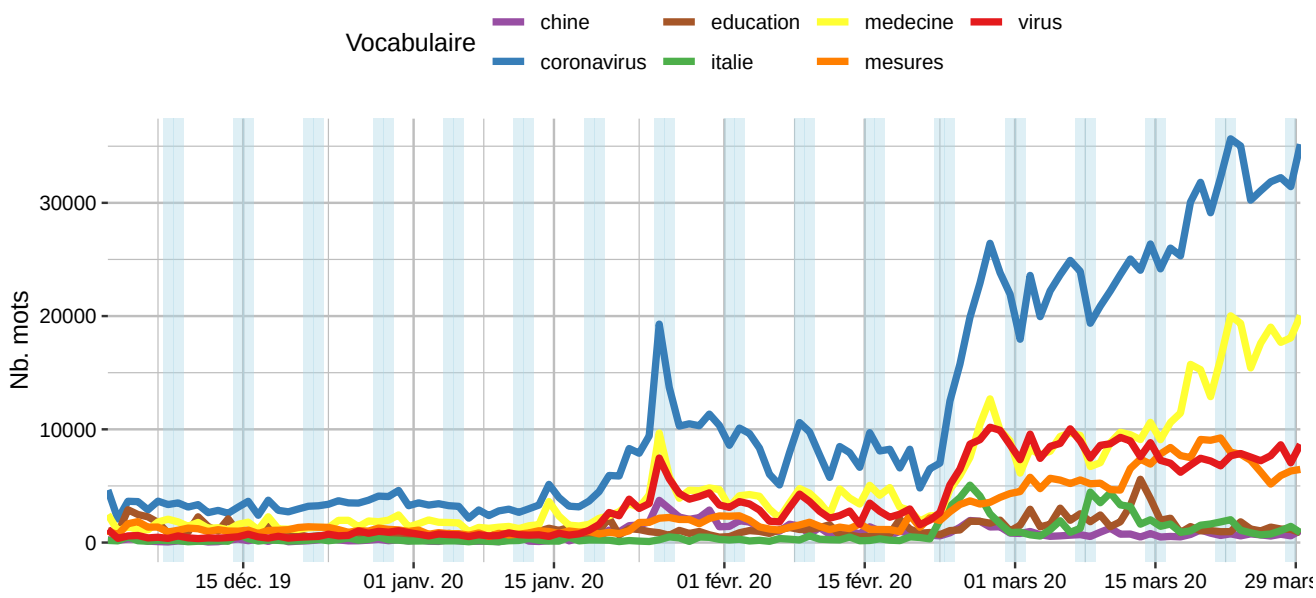


FIGURE 15 – Nombre d'occurrences des mots de chaque groupe de vocabulaire dans les transcriptions de l'ensemble des canaux d'information en continu.

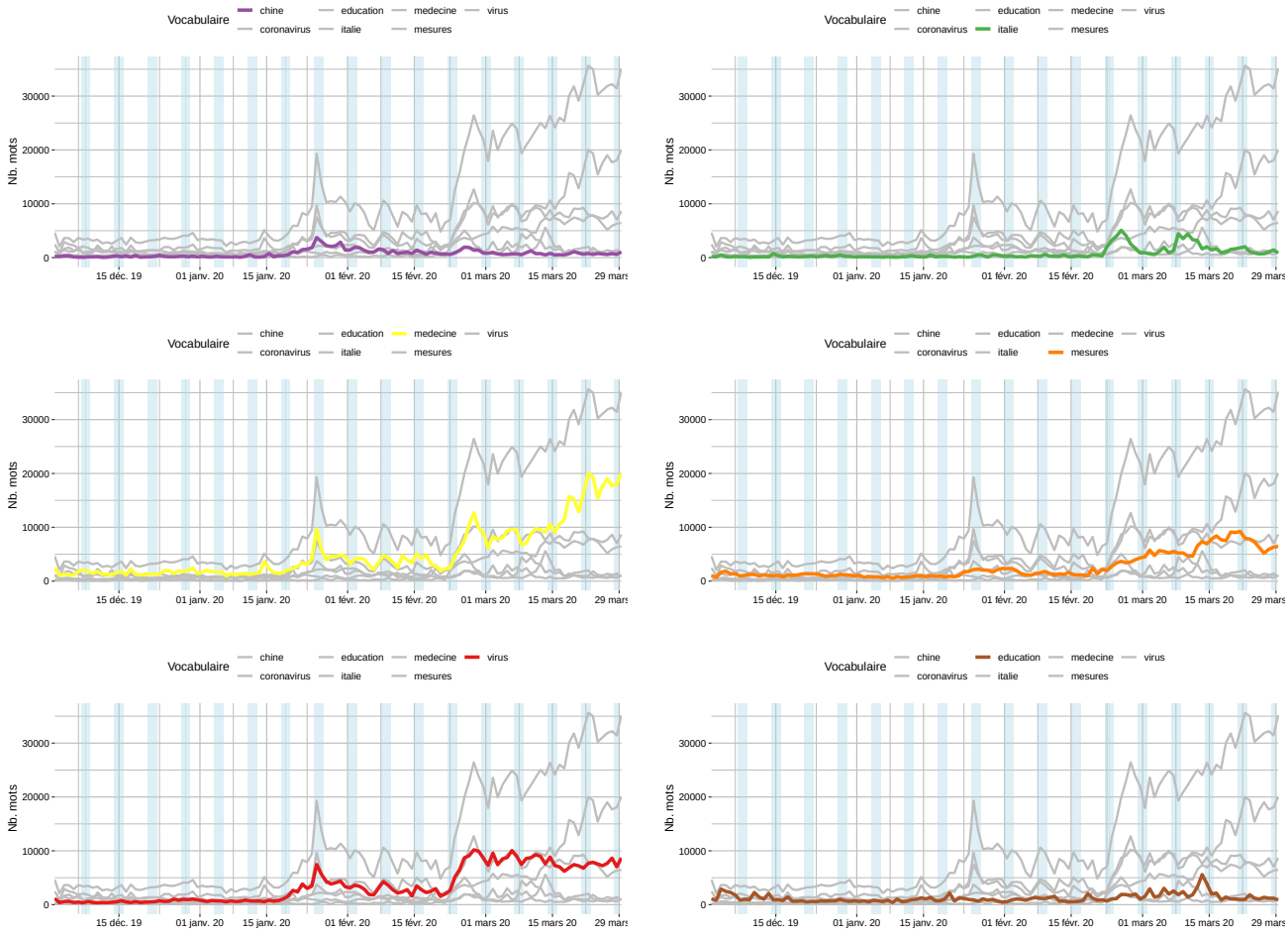
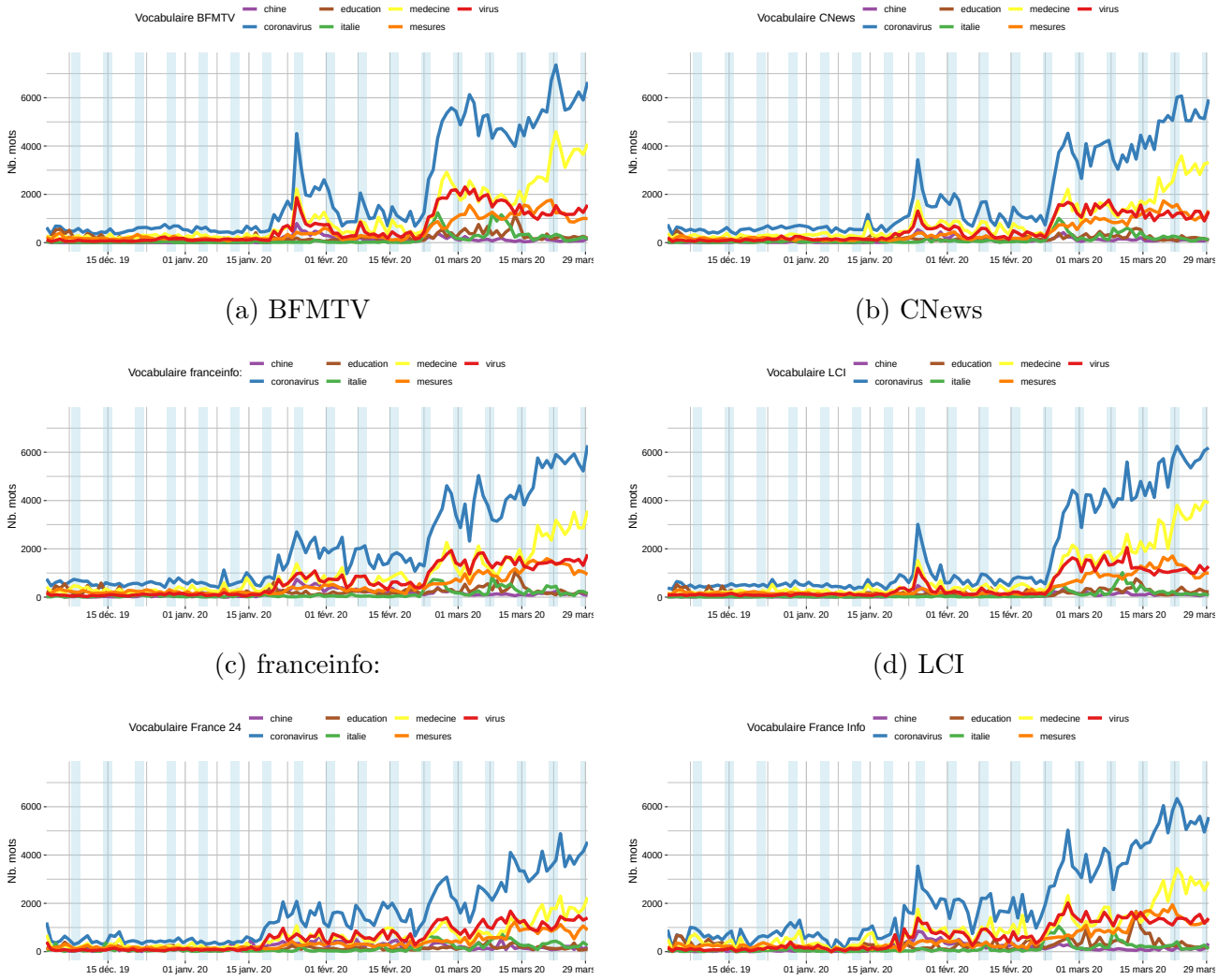


FIGURE 16 – Nombre d'occurrences des mots dans les transcriptions de l'ensemble des canaux d'information en continu pour chaque groupe de vocabulaire



(a) BFMTV

(b) CNews

(c) franceinfo:

(d) LCI

(e) France 24

(f) France Info

FIGURE 17 – Nombre d'occurrences des mots de chaque groupe de vocabulaire dans les transcriptions des canaux d'information en continu.

4.4 Les JT du soir et les matinales radio

L'idée est ici d'étudier la structure du JT et les choix de hiérarchisation de l'information. Sur la figure 18, on représente les jours en abscisse et les heures en ordonnée. On peut ainsi suivre l'évolution dans le temps (de gauche à droite) et au sein du JT (de bas en haut) des différentes thématiques liées au coronavirus. Chaque plage horaire est analysée avec une fenêtre glissante de 4 minutes [8] au sein de laquelle on compte le nombre d'occurrences des mots du vocabulaire. On normalise ensuite pas le maximum obtenu pour l'ensemble des chaînes. En annexe, les normalisations sont faites par chaîne et par groupe de mots pour mieux discerner les moments où certaines thématiques sont abordées.

La structuration temporelle apparaît bien pour les deux principaux JT de TF1 et France 2 et sont extrêmement similaires. Comme pour les autres médias on observe une première médiatisation en ouverture du JT la dernière semaine de janvier. Le sujet ne disparaît pas mais est traité plus tardivement ensuite (autour de 20h10). À partir de la dernière semaine de février, le sujet revient à la une jusqu'à prendre une place prépondérante puisqu'il est traité pendant 20 minutes. L'allongement de la durée des JT sur ces deux chaînes autour du 15 mars est également nettement visible. Nous avons tracer le profil du vocabulaire *coronavirus* de ces deux chaînes sur la figure 19. Saurez-vous les distinguer ?

Pour l'émission C dans l'air de France 5, on voit bien apparaître le caractère mono-thématique de chaque émission. De la même manière, la structure des matinales radio apparaît bien avec cette

8. avec un chevauchement de 2 minutes

visualisation du conducteur. Ainsi, notamment sur Europe 1 et France Inter, les créneaux de la matinale dédiées aux informations ressortent clairement sur la figure 20 puisqu'ils sont à heure fixe.

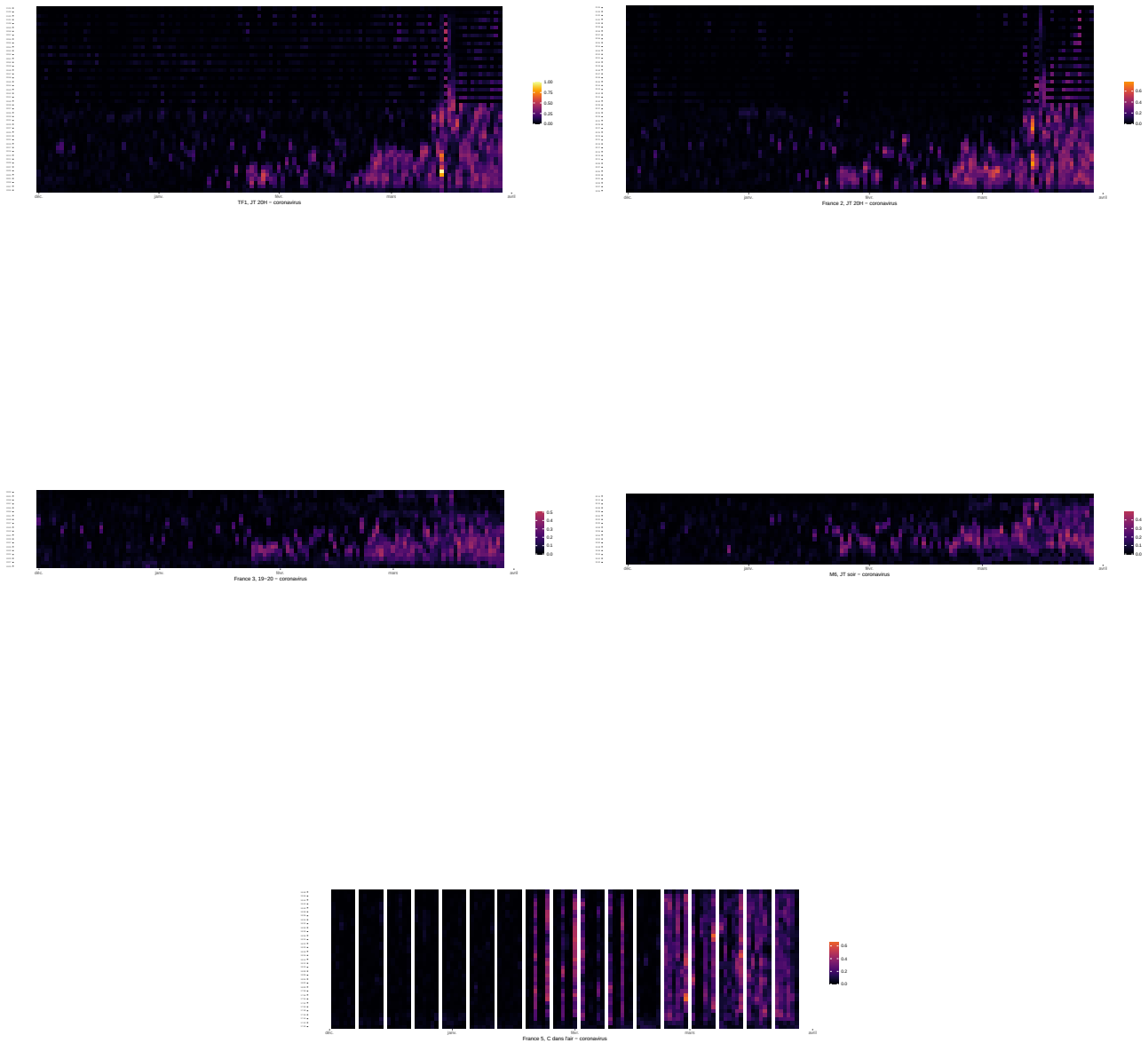


FIGURE 18 – Timelines pour toutes les chaînes TV, vocabulaire coronavirus (normalisation globale)

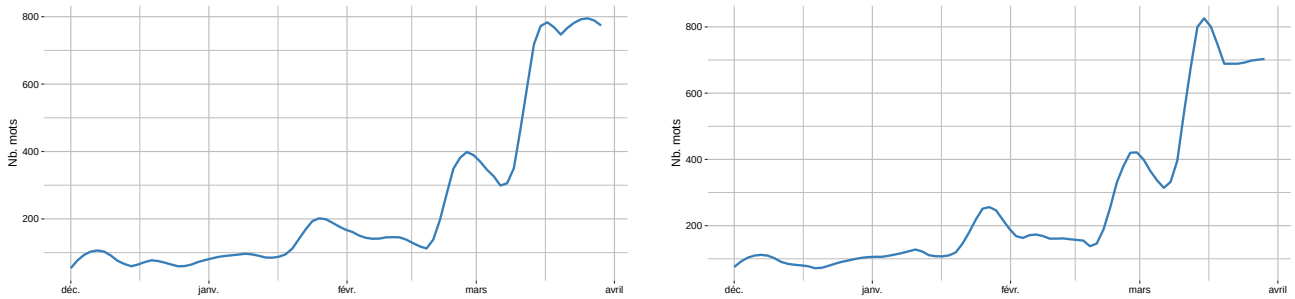


FIGURE 19 – Comparaison TF1 - France 2 sur le vocabulaire *coronavirus*

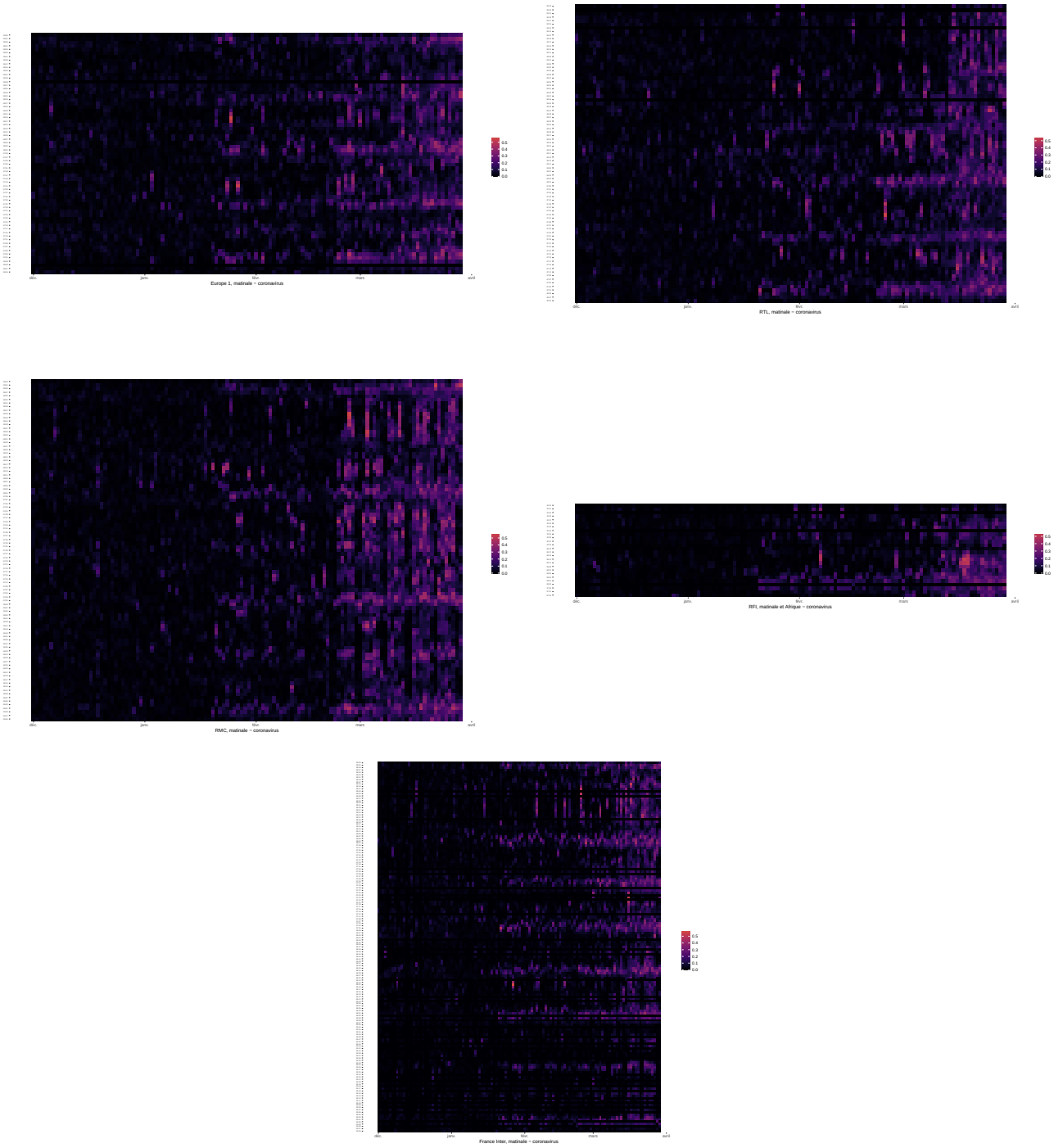


FIGURE 20 – Timelines pour toutes les radios, vocabulaire coronavirus (normalisation globale)

5 Résultats, chronologie et données externes

5.1 Comparaison de la médiatisation sur les différents médias

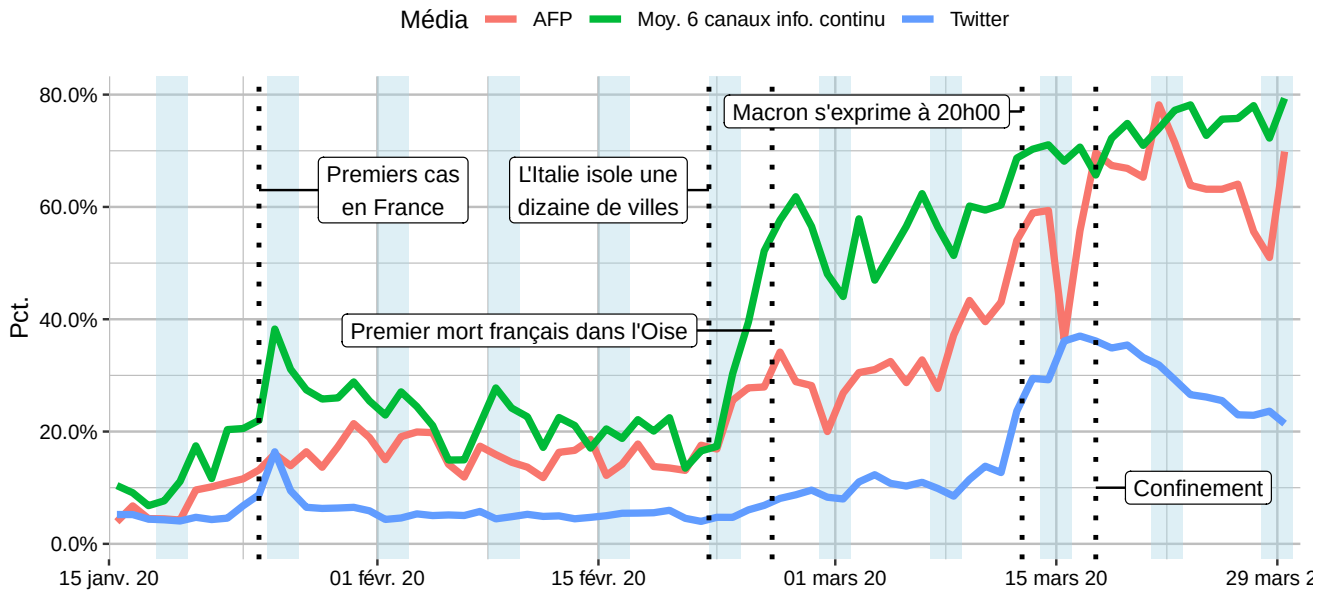


FIGURE 21 – Estimation du pourcentage des contenus diffusés qui sont liés aux principales thématiques du coronavirus : virus, médecine et mesures prises

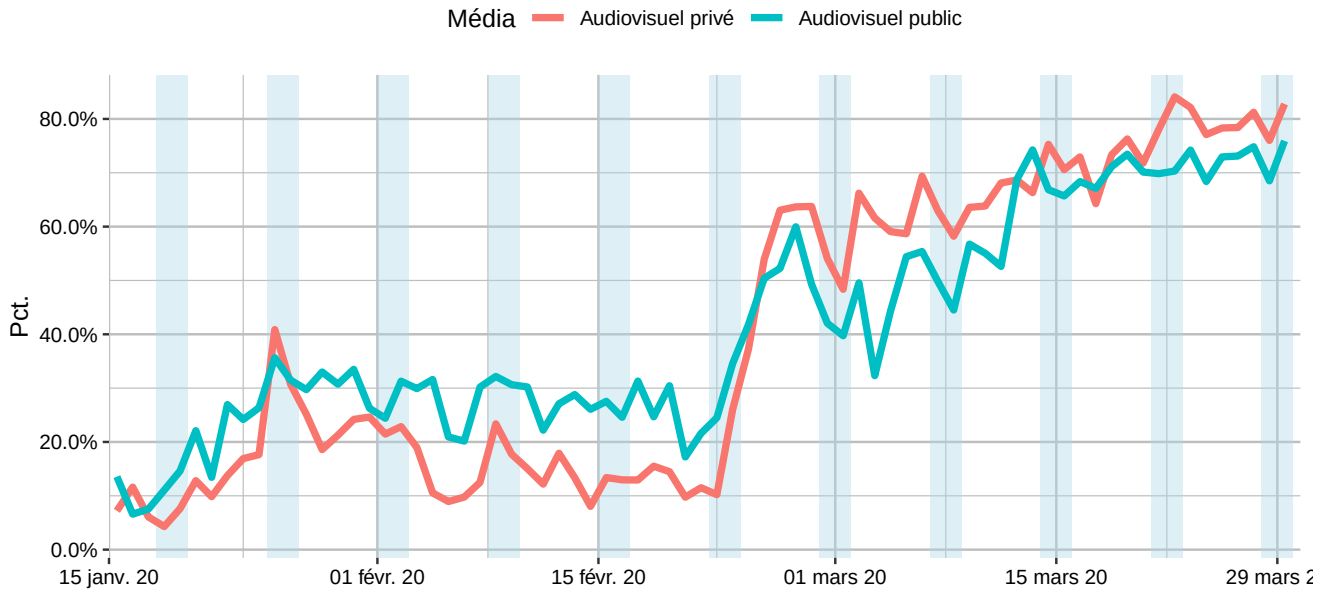


FIGURE 22 – Estimation du pourcentage des contenus diffusés qui sont liés aux principales thématiques du coronavirus entre les canaux publics (franceinfo:, France 24 et France Info) et privés (BFMTV, CNews et LCI)

5.2 Comparaison de la médiatisation et du nombre de cas du coronavirus

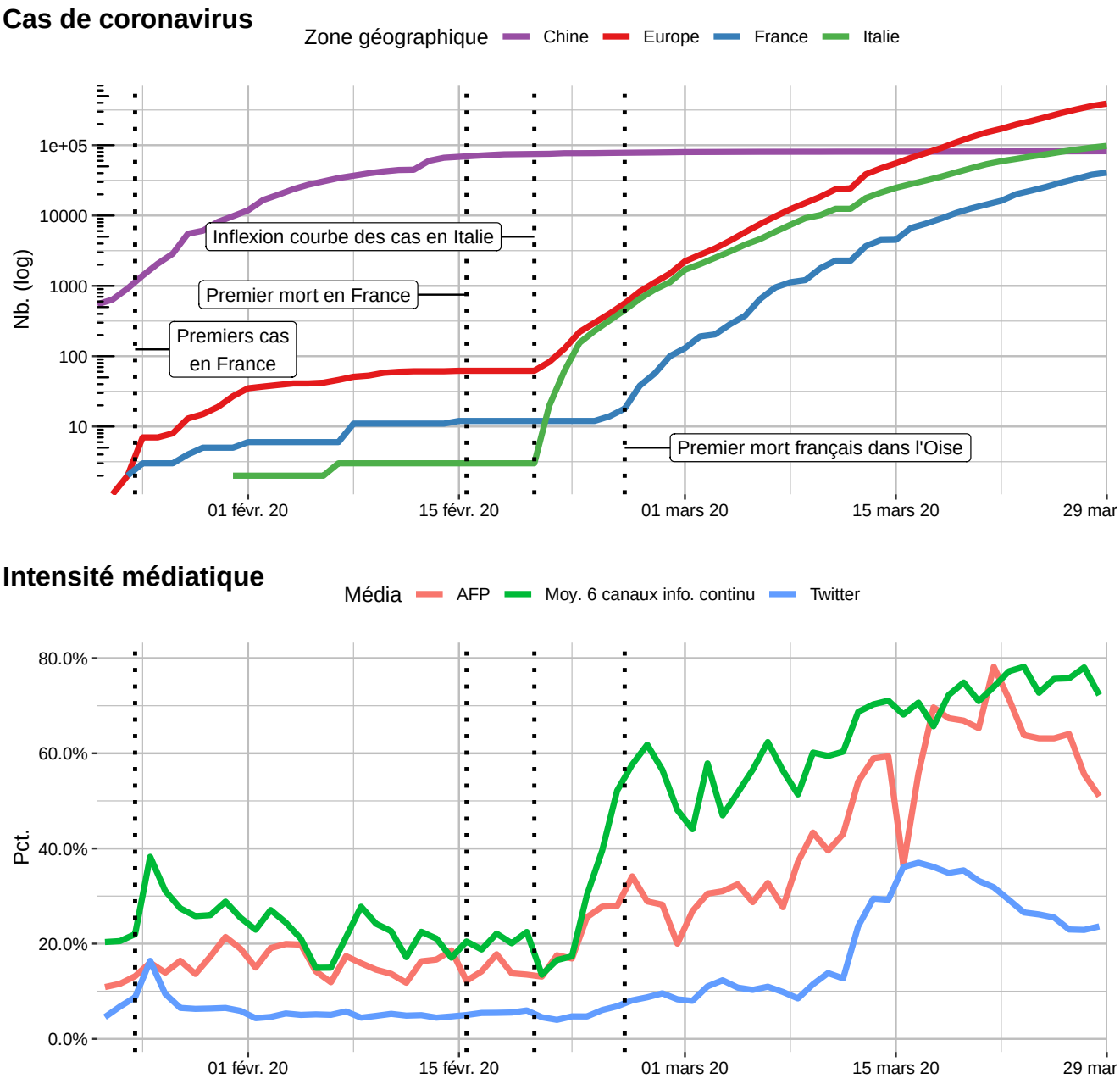


FIGURE 23 – Intensité médiatique comparée aux cas de coronavirus recensés.

5.3 Comparaison de la médiatisation et des cours de bourse

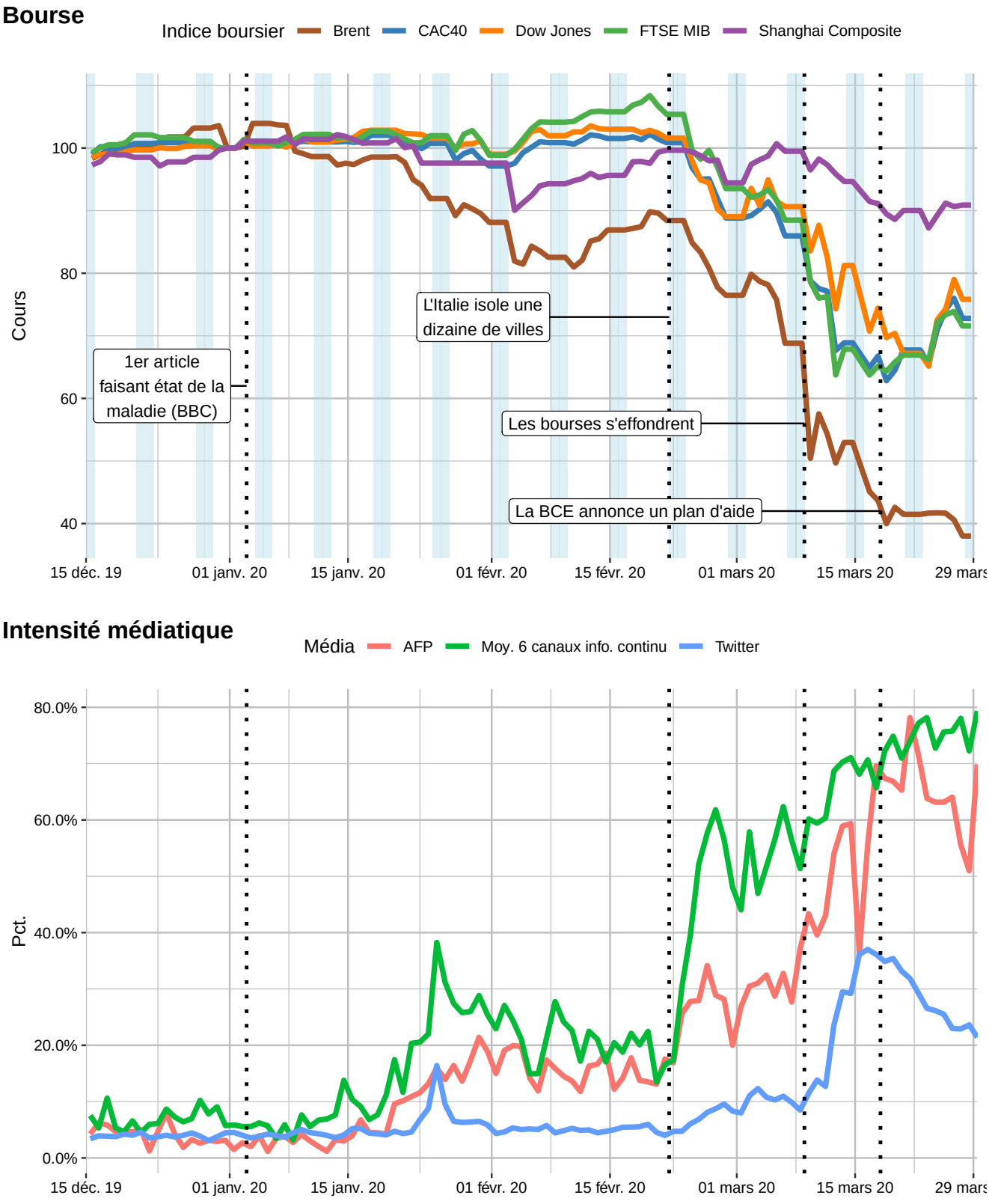


FIGURE 24 – Intensité médiatique du coronavirus comparé aux cours de bourse.

5.4 Étude de la médiatisation chloroquine / Didier Raoult

Nous utilisons un vocabulaire simplifié [9] pour regarder la médiatisation liée à la chloroquine. Contrairement aux autres quantifications effectuées sur les données Twitter, nous avons ici conservé les mentions afin de ne pas invisibiliser le compte Twitter @raoult_didier. Nous n'utilisons que l'approche textométrique. Ce vocabulaire a vocation à être intégré à celui de l'étude globale dans une prochaine version.

groupe	liste des préfixes
raoult	raoult, didierraoult, ihu
chloroquine	chloroquine, nivaquine, hydroxychloroquine*, plaquenil*

TABLE 8 – Vocabulaire chloroquine / Raoult - groupes de mots utilisés (préfixes)

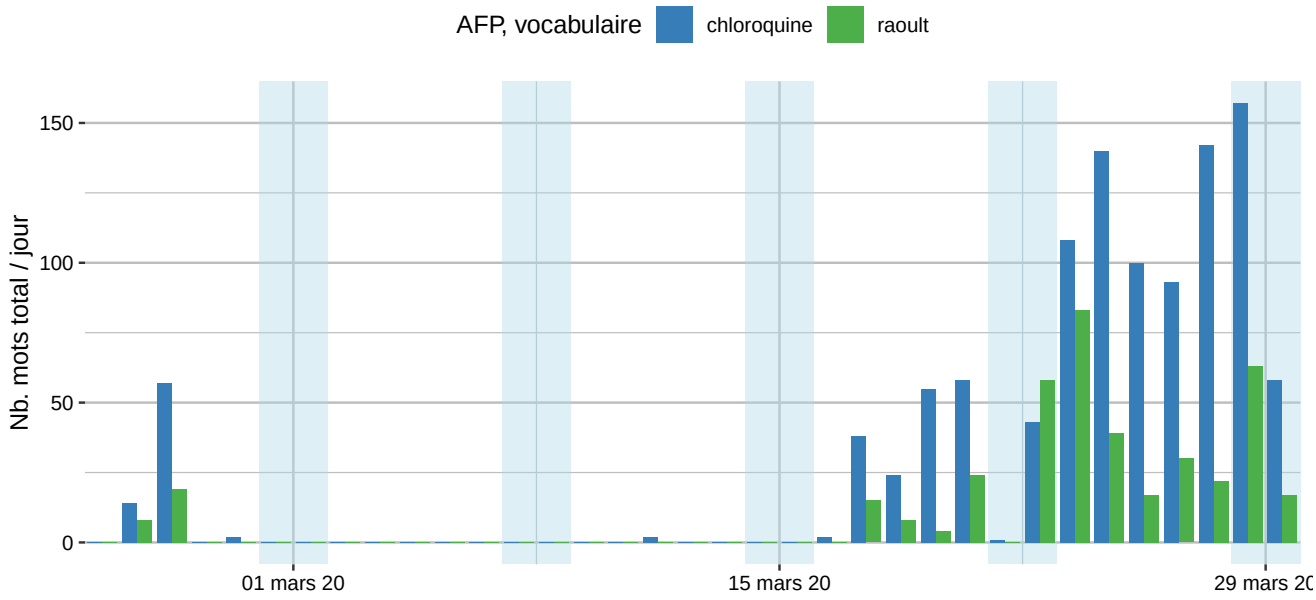


FIGURE 25 – AFP - chloroquine / Raoult

9. avec une étoile : mots non détectés par le logiciel de transcription, comptés uniquement sur l'AFP et Twitter

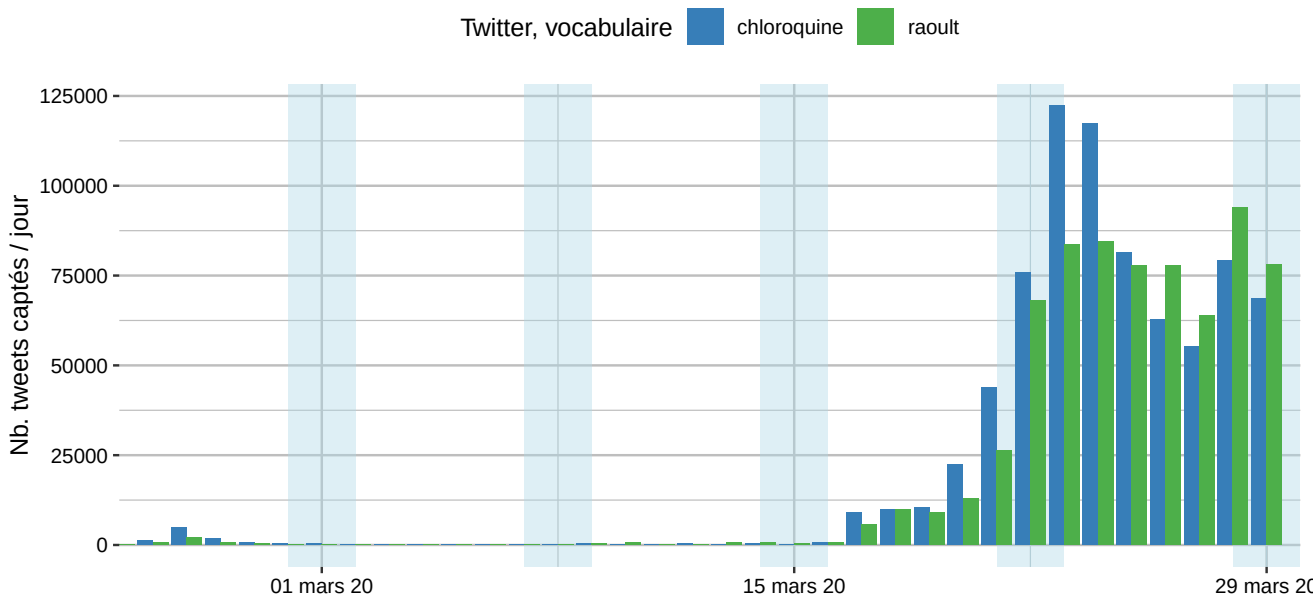


FIGURE 26 – Twitter - chloroquine / Raoult

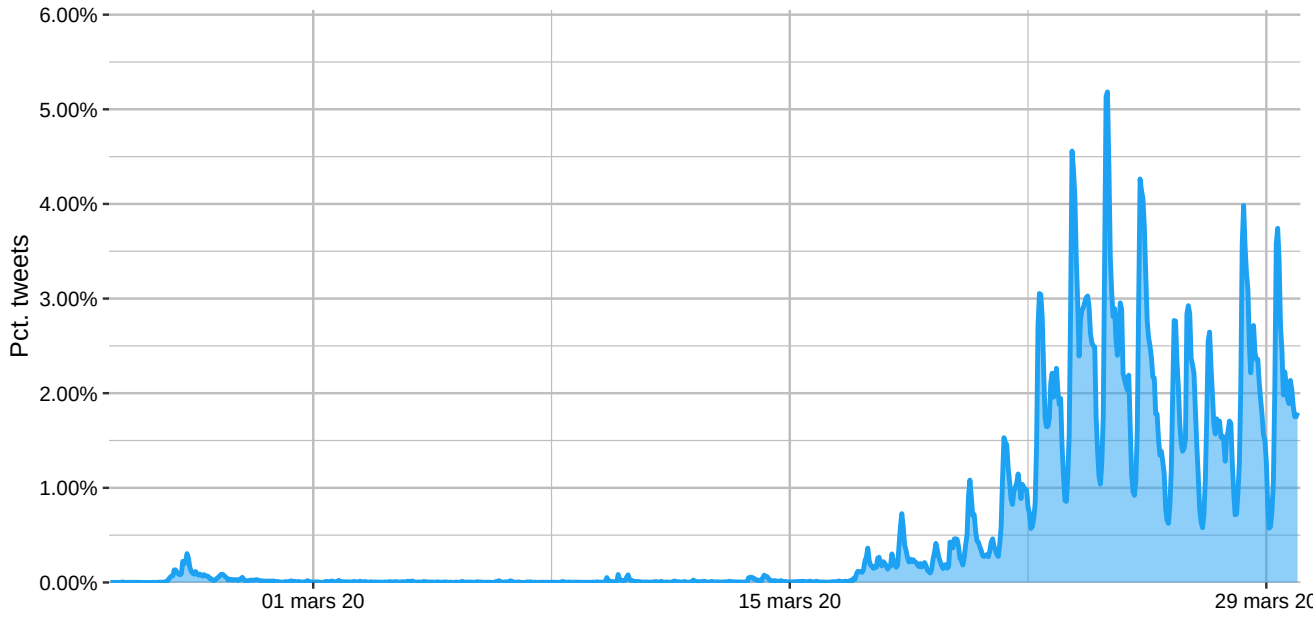
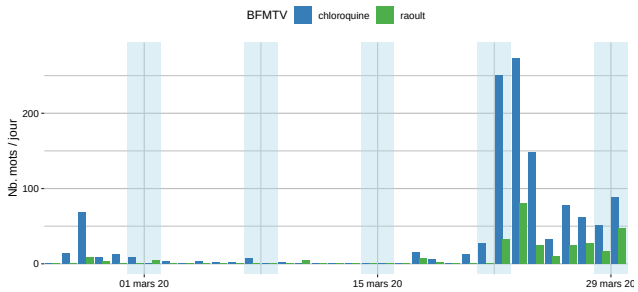
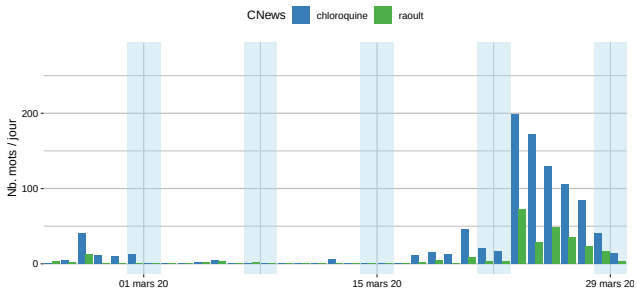


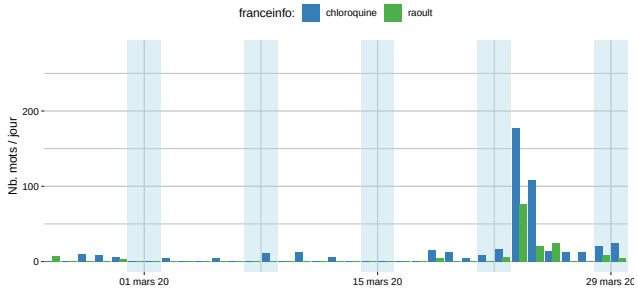
FIGURE 27 – Pourcentage de tweets liés à la chloroquine et/ou Didier Raoult



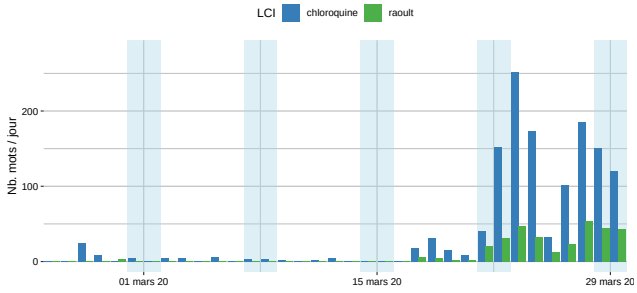
(a) BFMTV



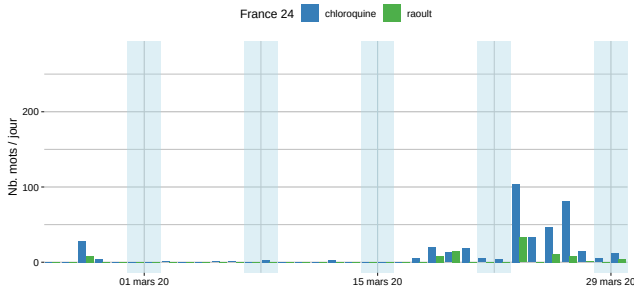
(b) CNews



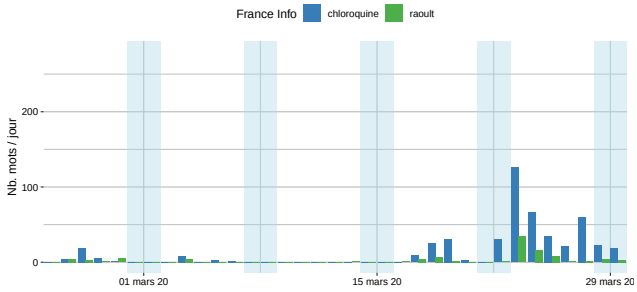
(c) franceinfo:



(d) LCI

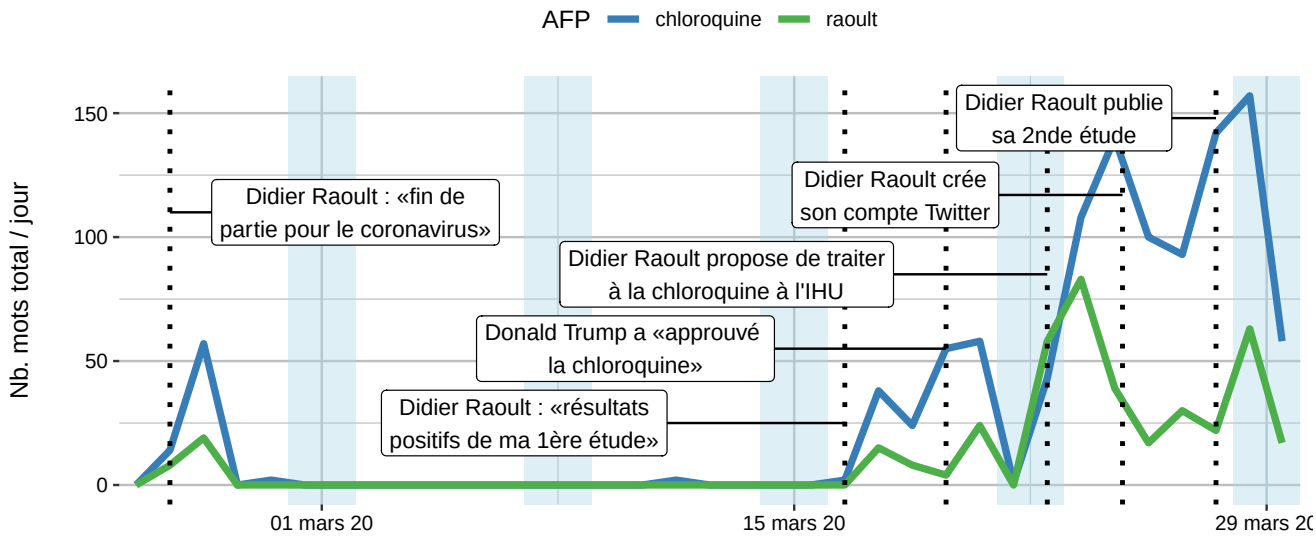


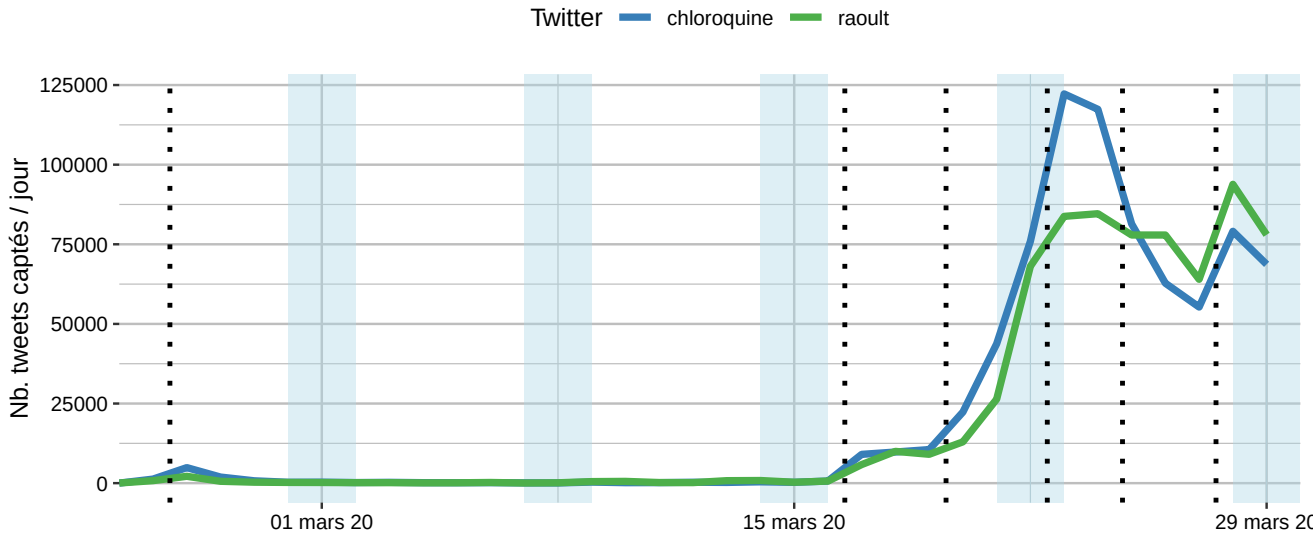
(e) France 24

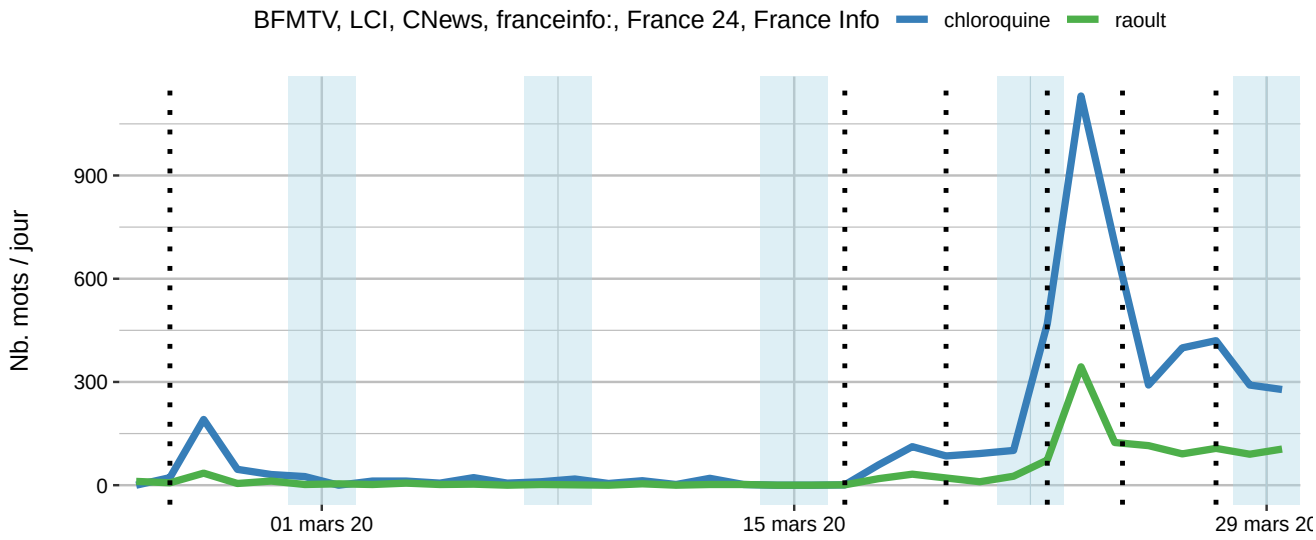


(f) France Info

FIGURE 28 – Canaux d'information en continu - chloroquine / Raoult


AFP
chloroquine
raoult
Nb. mots total / jour
150
100
50
0
Didier Raoult : «fin de partie pour le coronavirus»
Didier Raoult propose de traiter à la chloroquine à l'IHU
Donald Trump a «approuvé la chloroquine»
Didier Raoult : «résultats positifs de ma 1ère étude»
Didier Raoult crée son compte Twitter
Didier Raoult publie sa 2nde étude
01 mars 20
15 mars 20
29 mars 20


Twitter
chloroquine
raoult
Nb. tweets captés / jour
125000
100000
75000
50000
25000
0
01 mars 20
15 mars 20
29 mars 20


BFMTV, LCI, CNews, franceinfo:, France 24, France Info
chloroquine
raoult
Nb. mots / jour
900
600
300
0
01 mars 20
15 mars 20
29 mars 20

6 Conclusion

Les principaux points à noter à ce stade de la médiatisation de l'épidémie sont :

— la particularité de cette étude est qu'elle quantifie la proportion de la production d'information qui est liée au coronavirus et non simplement une quantité d'information
— une saturation de l'espace médiatique : depuis le début du confinement, canaux d'information en continu et AFP frôlent avec les 80% de contenus publiés liés au coronavirus (figure 21, page 20). N'y a-t-il plus d'actualité en dehors du coronavirus à cause du confinement et de la mise à l'arrêt de la vie sociale ou bien n'y a-t-il plus d'espace médiatique disponible pour en parler ?
— les canaux d'information en continu arrivent beaucoup plus vite à ce seuil que l'AFP
— les deux principales raisons que nous notons pour une nette augmentation de la médiatisation sont l'apparition du premier cas en France et la décision de l'Italie de mettre une dizaine de villes sous cloches en Lombardie suivie de peu par le premier mort français.
— sur Twitter, le vrai pic démarre la veille de la première intervention télévisés d'Emmanuel Macron. On y observe une légère décrue de la thématique depuis le dimanche du premier tour des élections municipales
— les JT de TF1 et France 2 se ressemblent très fortement (figure 18, page 18)
— le premier mort en France ne provoque, médiatiquement, rien de particulier

Ce travail est encore en cours et de nombreuses pistes de travail sont envisagées pour le compléter. L'augmentation du périmètre est un premier axe, avec l'ajout de la radio et de la presse et la complétion au fur et à mesure des jeux de données existants. Concernant les approches quantitatives, nous prévoyons d'utiliser des approches algorithmique pour déterminer les vocabulaires (*Topic Modeling*, *clustering*, . . .).

7 Remerciements

Cette étude n'aurait pas été possible sans la disponibilité des différents jeux de données utilisés. Nous tenons à chaleureusement remercier Béatrice Mazoyer pour sa solution de captation des tweets qui fonctionne sans discontinuer depuis l'été 2018. La captation des flux télé et radio fait partie des missions de l'Ina au titre du dépôt légal. Ces données sont rapidement disponibles pour effectuer des analyses, notamment grâce au travail de Thomas Drugeon et son logiciel Collgate. La transcription de ces flux ne serait pas possible sans le logiciel fourni par le LIUM [10] ni le travail de Pierre Letessier et Denis Rakulan qui l'ont intégré sur nos serveurs et ont développé la solution de pilotage. L'aide de Richard Poirot a été précieuse lors de la première étude [11], il a en effet eu la lourde et fastidieuse tâche de création de la vérité terrain qui a permis de calibrer l'algorithme de segmentation pour le calcul du temps d'antenne. Enfin, nous sommes reconnaissant à l'AFP de nous permettre d'avoir accès à ses dépêches via son API [11] grâce notamment à Denis Teyssou. L'ensemble des logiciels de captation et de transcription sont automatisés et stables, permettant ainsi une étude en période de confinement et en temps réel. Merci enfin à Agnès Saulnier, Béatrice Mazoyer, Éléonore Alquier, Antoine Bayet et Boris Jamet-Fournier pour leurs relecture, commentaires et suggestions pertinentes.

Références

[1] Emily Chen, Kristina Lerman, and Emilio Ferrara. Covid-19 : The first public coronavirus twitter dataset, 2020. [Voir en ligne].

[2] Johns Hopkins CSSE. 2019 novel coronavirus covid-19 (2019-ncov) data repository, 2020. [Voir en ligne].

[3] Xavier Demagny. Youtube, twitter, facebook : Didier raoult est devenu une star du web (et pas que pour le meilleur). 2020. [Voir en ligne].

[4] Nicolas Hervé. OTMedia, the TransMedia News Observatory. In *FIAT/IFTA Media Management Seminar 2019*, Stockholm, Sweden, May 2019. [Voir en ligne].

10. https://lium.univ-lemans.fr/
11. https://developers.afp.com/fr

[5] Nicolas Hervé, Pierre Letessier, Mathieu Derval, and Hakim Nabi. Amalia.js : An open-source metadata driven html5 multimedia player. In *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference*, MM '15, pages 709–712, New York, NY, USA, 2015. ACM. [Voir en ligne].

[6] Stéphane Jourdain. Selon une étude, 94% des commentaires facebook portent sur le coronavirus. 2020. [Voir en ligne].

[7] Juliette Labracherie and Nicolas Hervé. Incendie de l'usine lubrizol à rouen et mort de jacques chirac : comment les chaînes info ont traité d'une double actualité. *La Revue des Médias*, 2019. [Voir en ligne].

[8] Béatrice Mazoyer, Julia Cagé, Céline Hudelot, and Marie-Luce Viaud. Real-time collection of reliable and representative tweets datasets related to news events. In *First International Workshop on Analysis of Broad Dynamic Topics over Social Media (BroDyn 2018) co-located with the 40th European Conference on Information Retrieval (ECIR 2018)*, Grenoble, France, March 2018. [Voir en ligne].

[9] Thomas Moysan. Mathilde guinaudeau (ipsos) : « twitter concentre 75conversations en ligne sur le coronavirus en france ». 2020. [Voir en ligne].

[10] Cyril Petit. Coronavirus : près de 19.000 articles chaque jour dans la presse française, un record. 2020. [Voir en ligne].

[11] Richard Poirot and Nicolas Hervé. Les « gilets jaunes », trou noir médiatique. *La Revue des Médias*, 2019. [Voir en ligne].

[12] Jérémie Rappaz, François Quellec, and Paul Ronga. Covid-19 : histoire d'une médiatisation. 2020. [Voir en ligne].

[13] Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28 :11–21, 1972. [Voir en ligne].

[14] Chengjun Sun, Wei Yang, Julien Arino, and Kamran Khan. Effect of media-induced social distancing on disease transmission in a two patch setting. *Mathematical Biosciences*, 230(2) :87 – 95, 2011.

[15] Natalia Tomashenko, Kévin Vythelingum, Anthony Rousseau, and Yannick Estève. LIUM ASR systems for the 2016 Multi-Genre Broadcast Arabic Challenge. In *IEEE Workshop on Spoken Language Technology*, San Diego, CA, USA, United States, December 2016. [Voir en ligne].

[16] Qingchu Wu, Xinchu Fu, Michael Small, and Xin-Jian Xu. The impact of awareness on epidemic spreading in networks. *Chaos : an interdisciplinary journal of nonlinear science*, 22(1) :013101, 2012.

8 Annexes

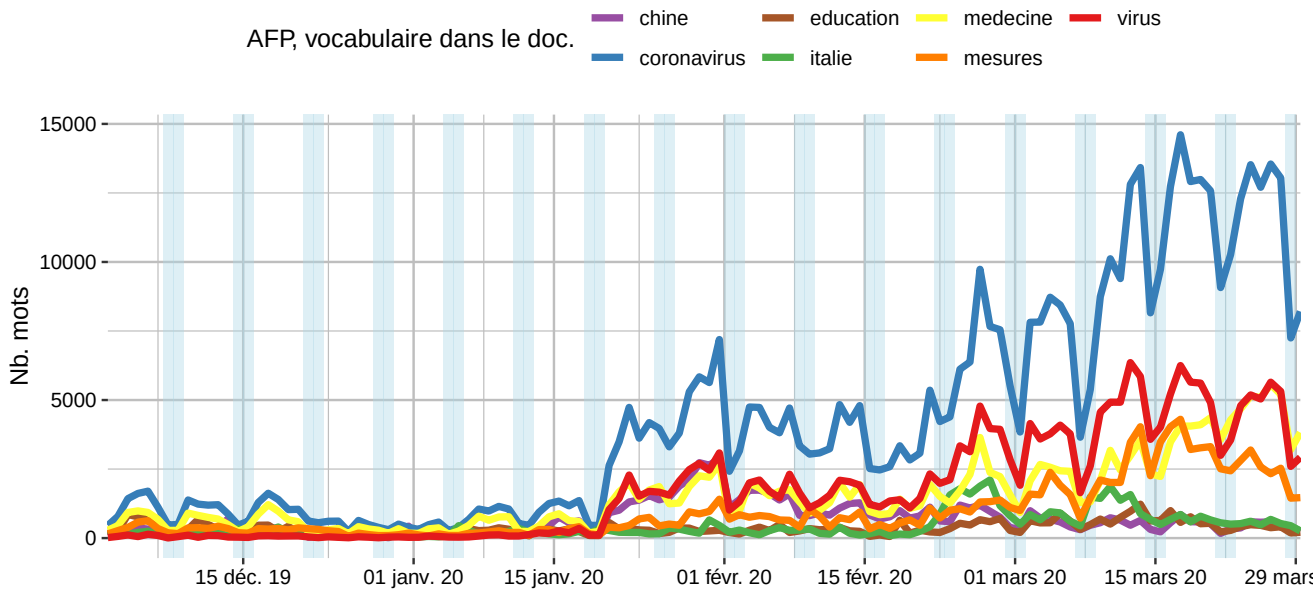


FIGURE 29 – Nombre d'occurrences des mots de chaque groupe de vocabulaire dans les dépêches AFP.

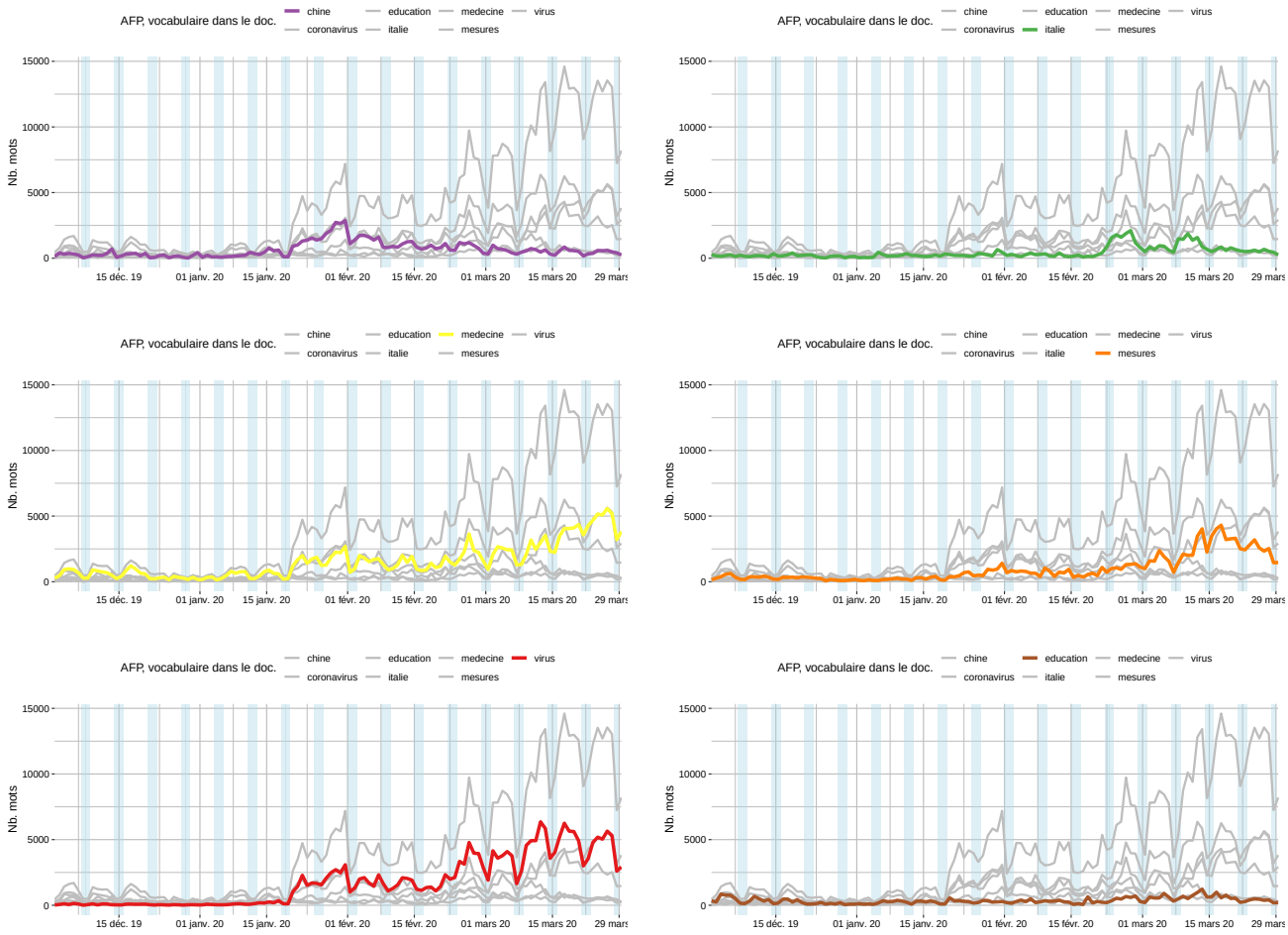


FIGURE 30 – Nombre d'occurrences des mots dans les dépêches AFP pour chaque groupe de vocabulaire

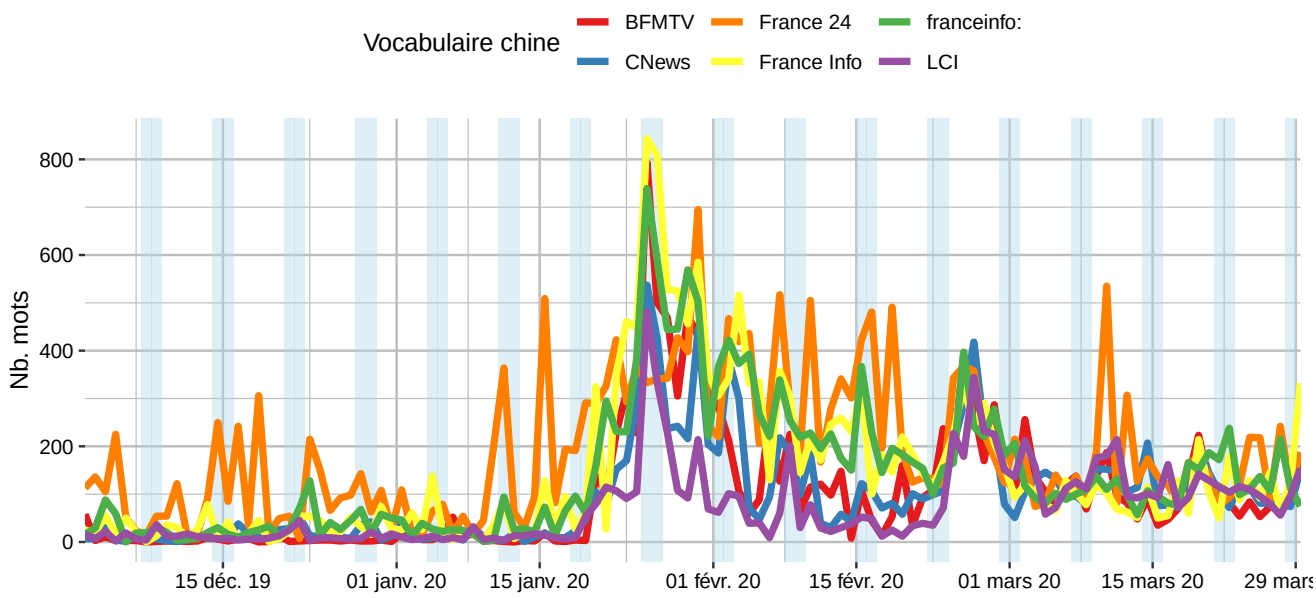


FIGURE 31 – Nombre d'occurrences des mots du groupe *chine* dans les transcriptions de l'ensemble des canaux d'information en continu

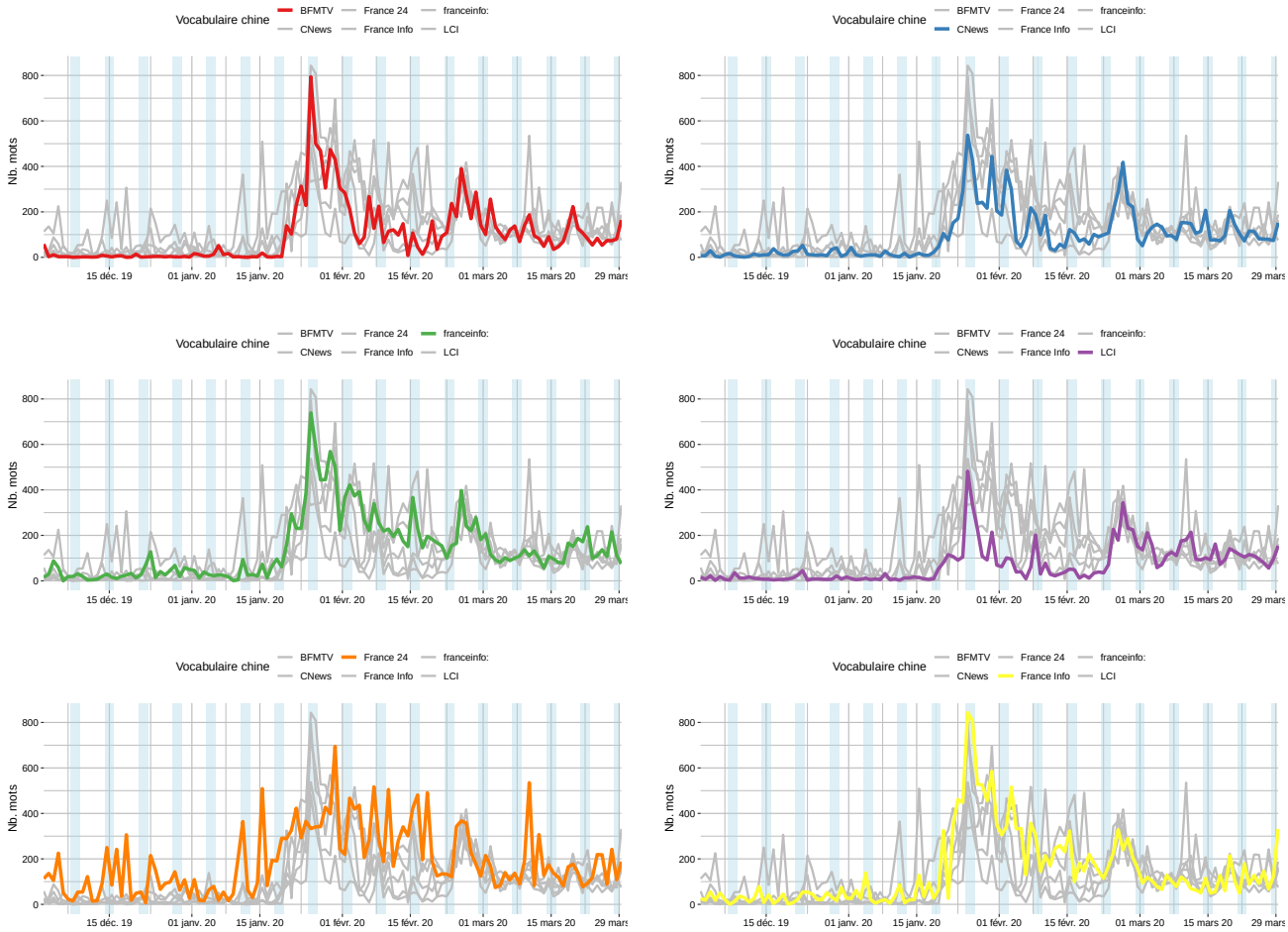


FIGURE 32 – Nombre d'occurrences des mots du groupe *chine* dans les transcriptions de chaque canal d'information en continu

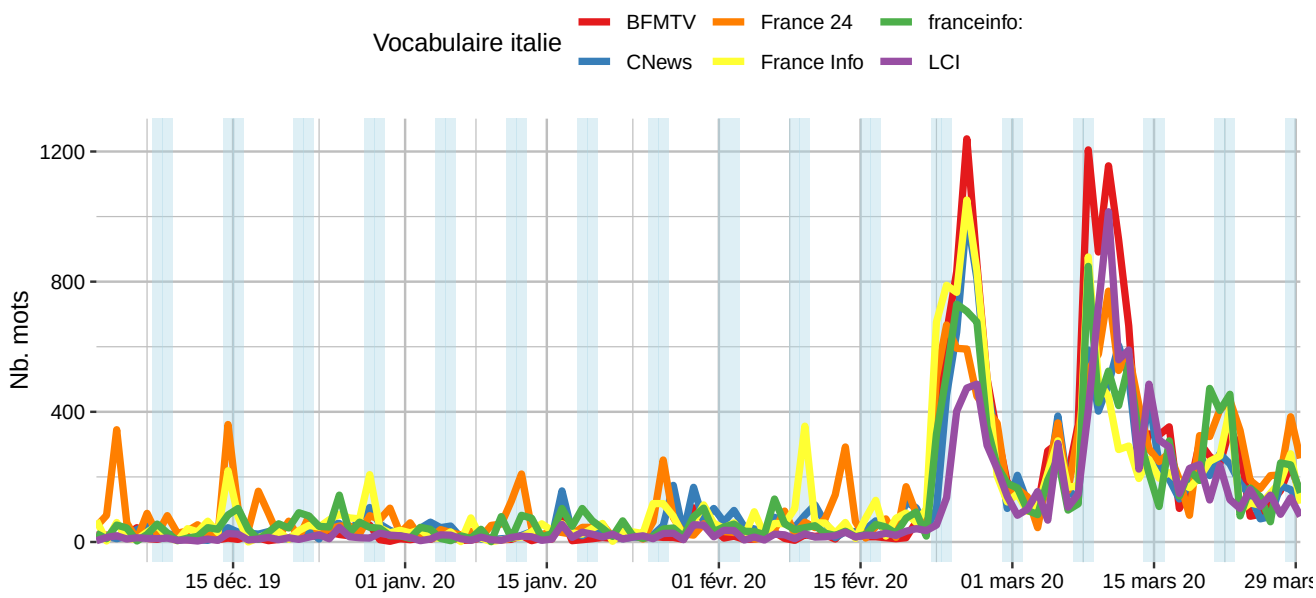


FIGURE 33 – Nombre d'occurrences des mots du groupe *italie* dans les transcriptions de l'ensemble des canaux d'information en continu

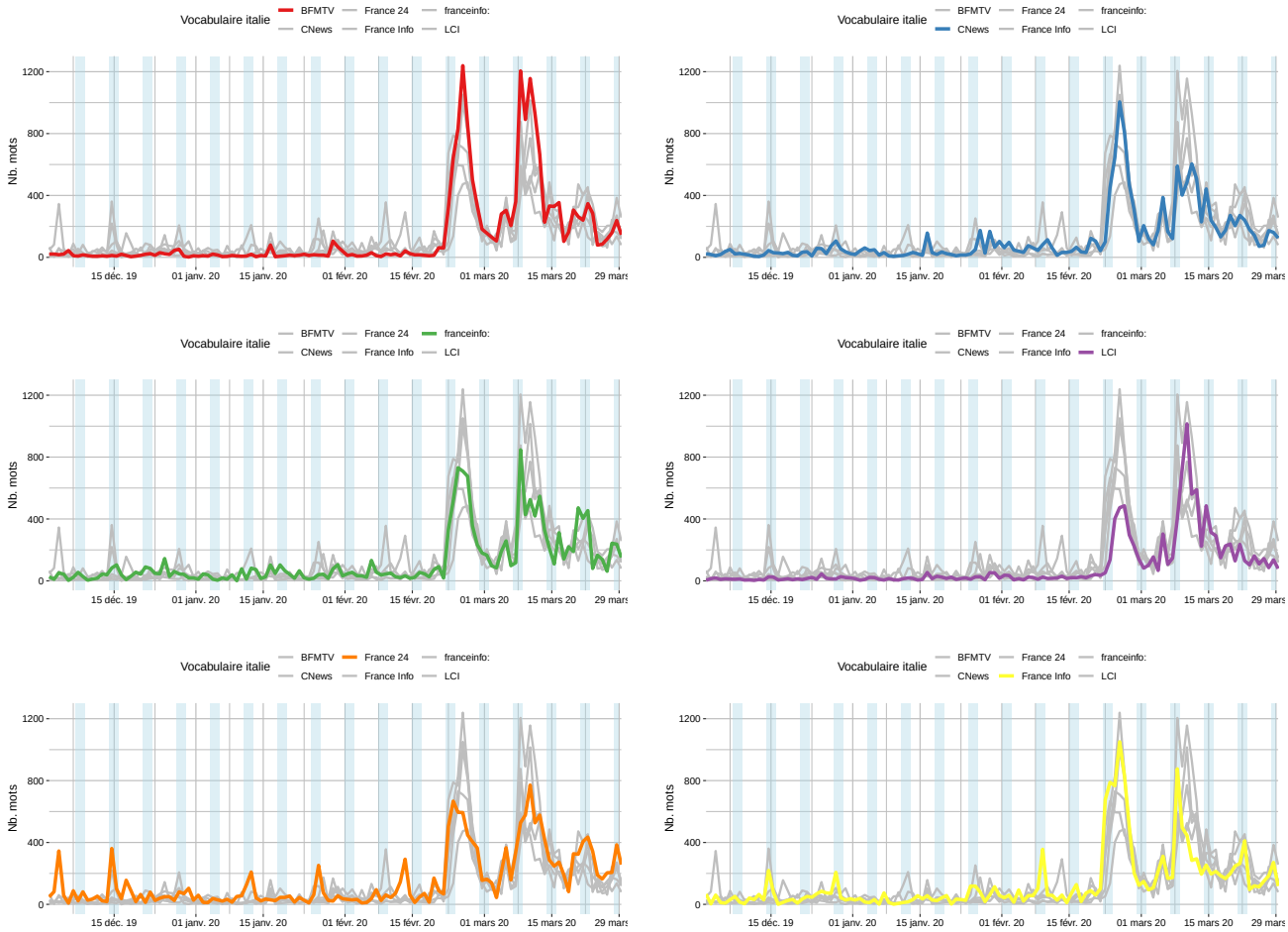


FIGURE 34 – Nombre d'occurrences des mots du groupe *italie* dans les transcriptions de chaque canal d'information en continu

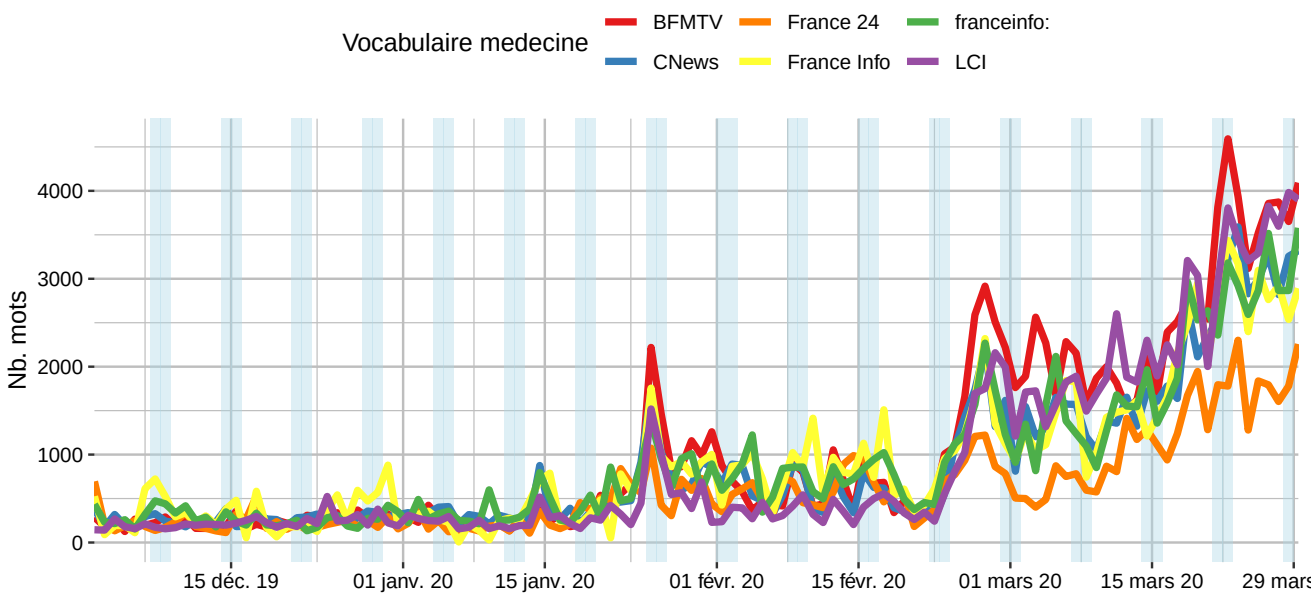


FIGURE 35 – Nombre d'occurrences des mots du groupe *medecine* dans les transcriptions de l'ensemble des canaux d'information en continu

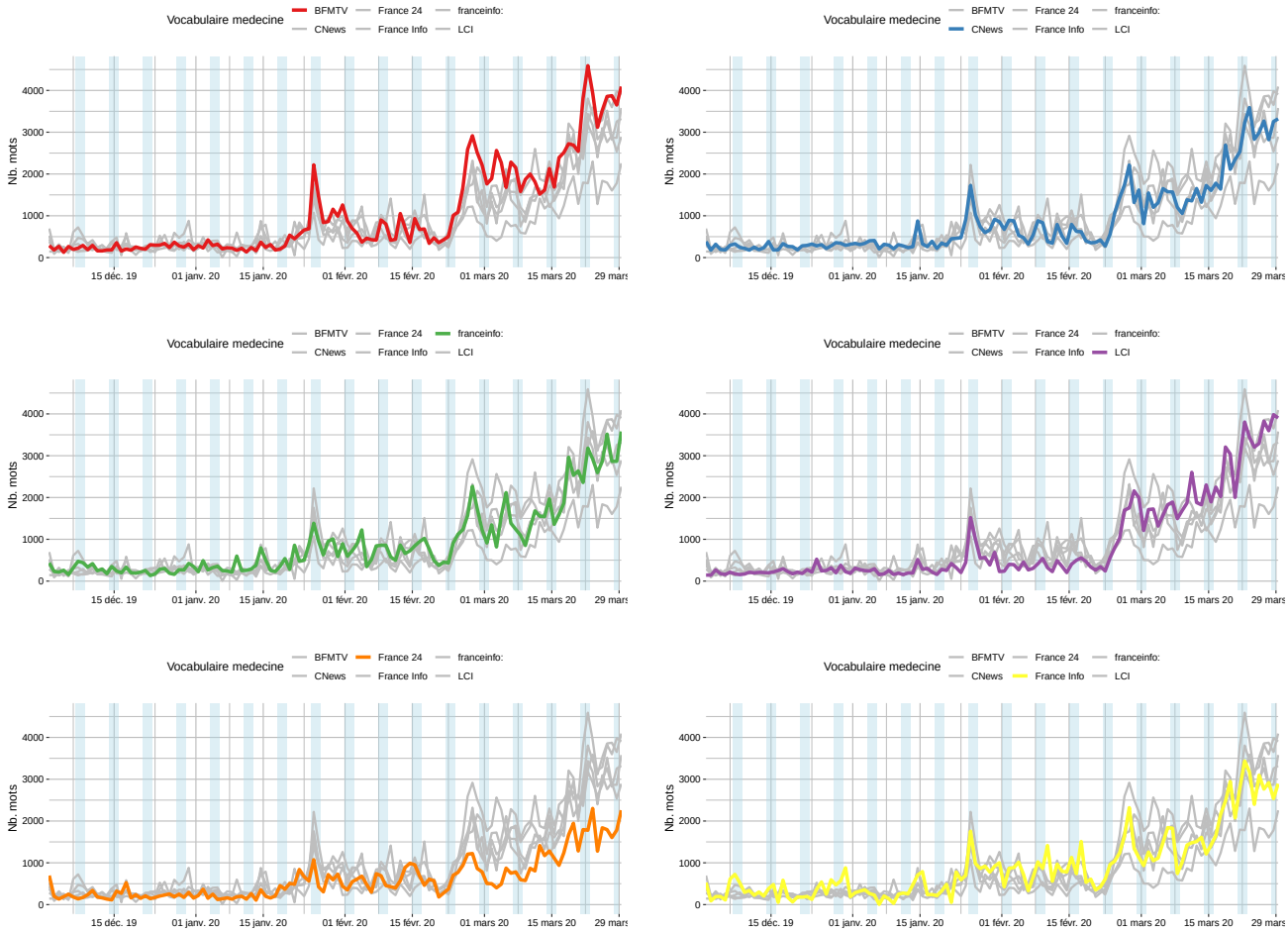


FIGURE 36 – Nombre d'occurrences des mots du groupe *medecine* dans les transcriptions de chaque canal d'information en continu

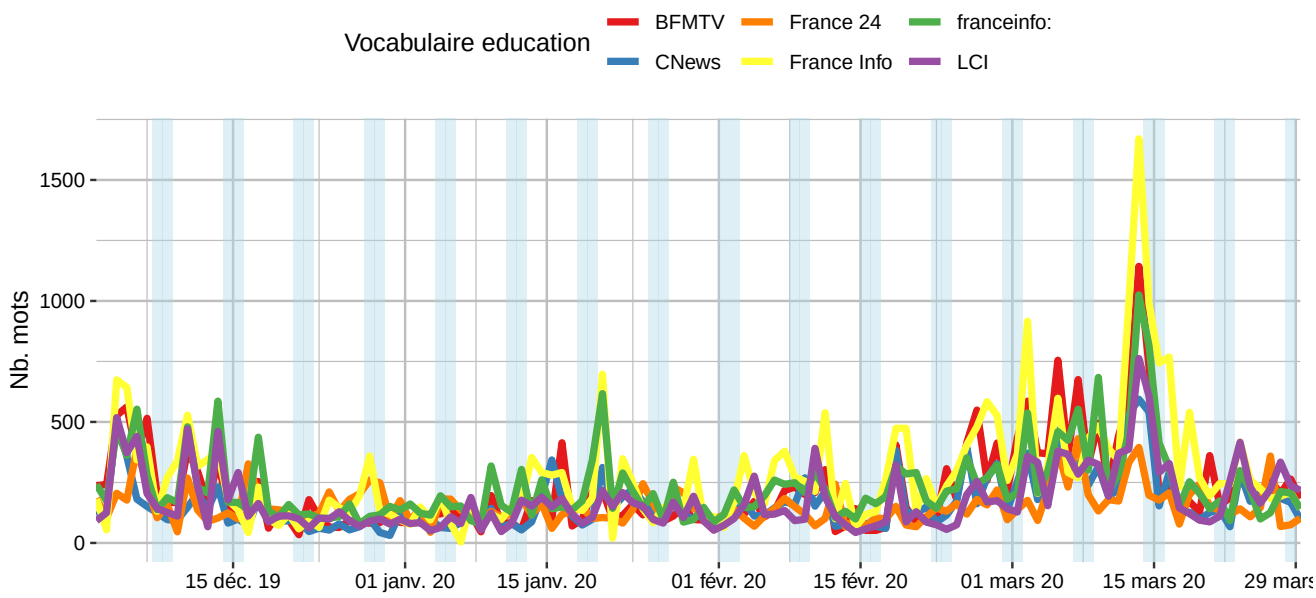


FIGURE 37 – Nombre d'occurrences des mots du groupe *education* dans les transcriptions de l'ensemble des canaux d'information en continu

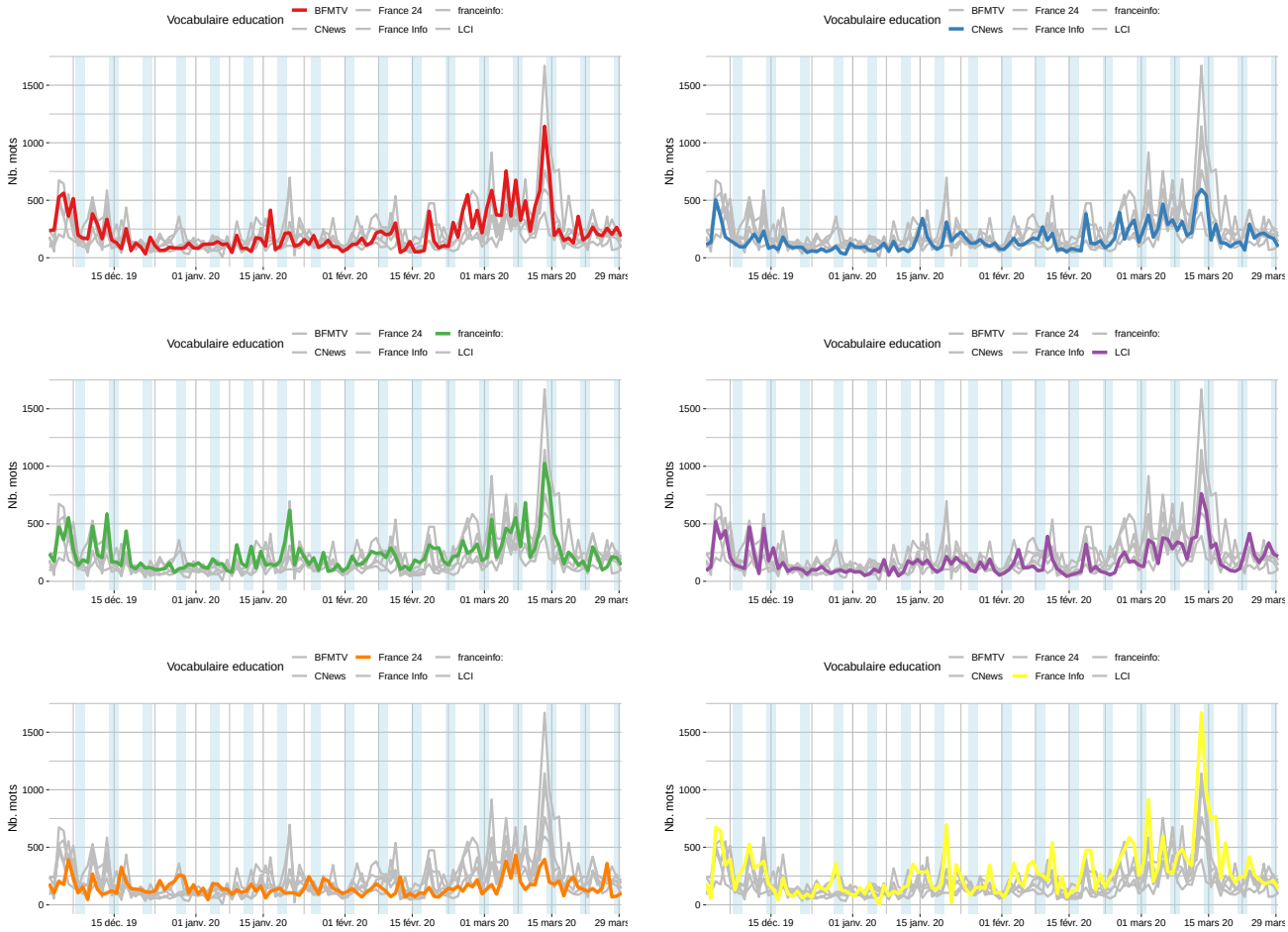


FIGURE 38 – Nombre d'occurrences des mots du groupe *education* dans les transcriptions de chaque canal d'information en continu

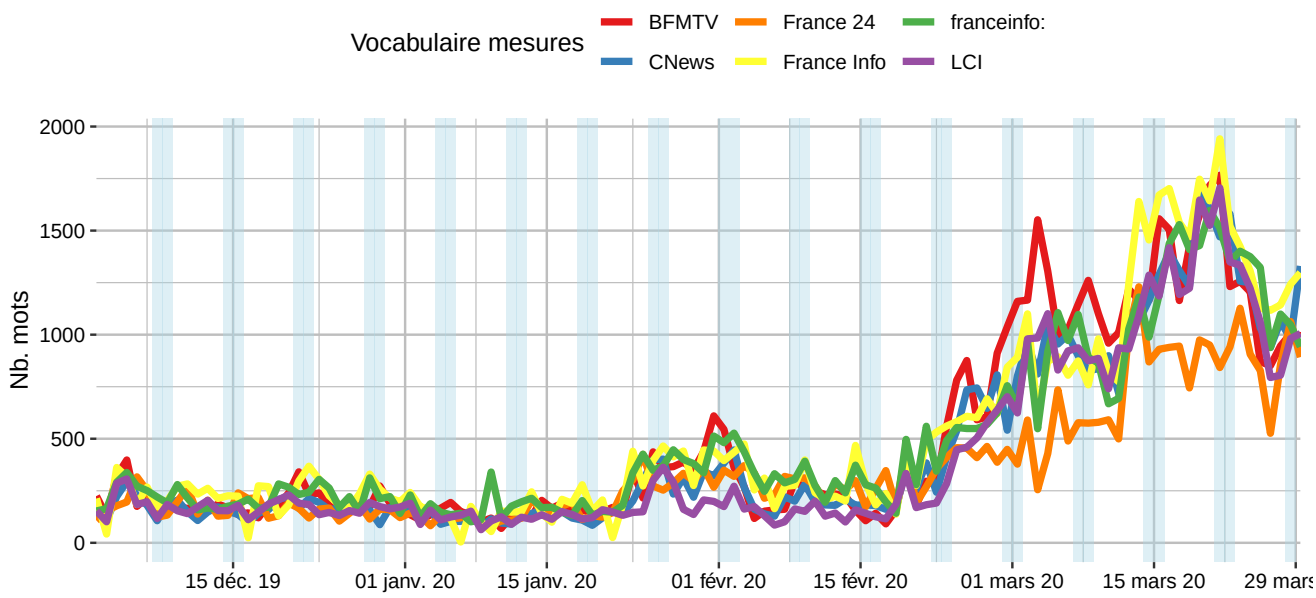


FIGURE 39 – Nombre d'occurrences des mots du groupe *mesures* dans les transcriptions de l'ensemble des canaux d'information en continu

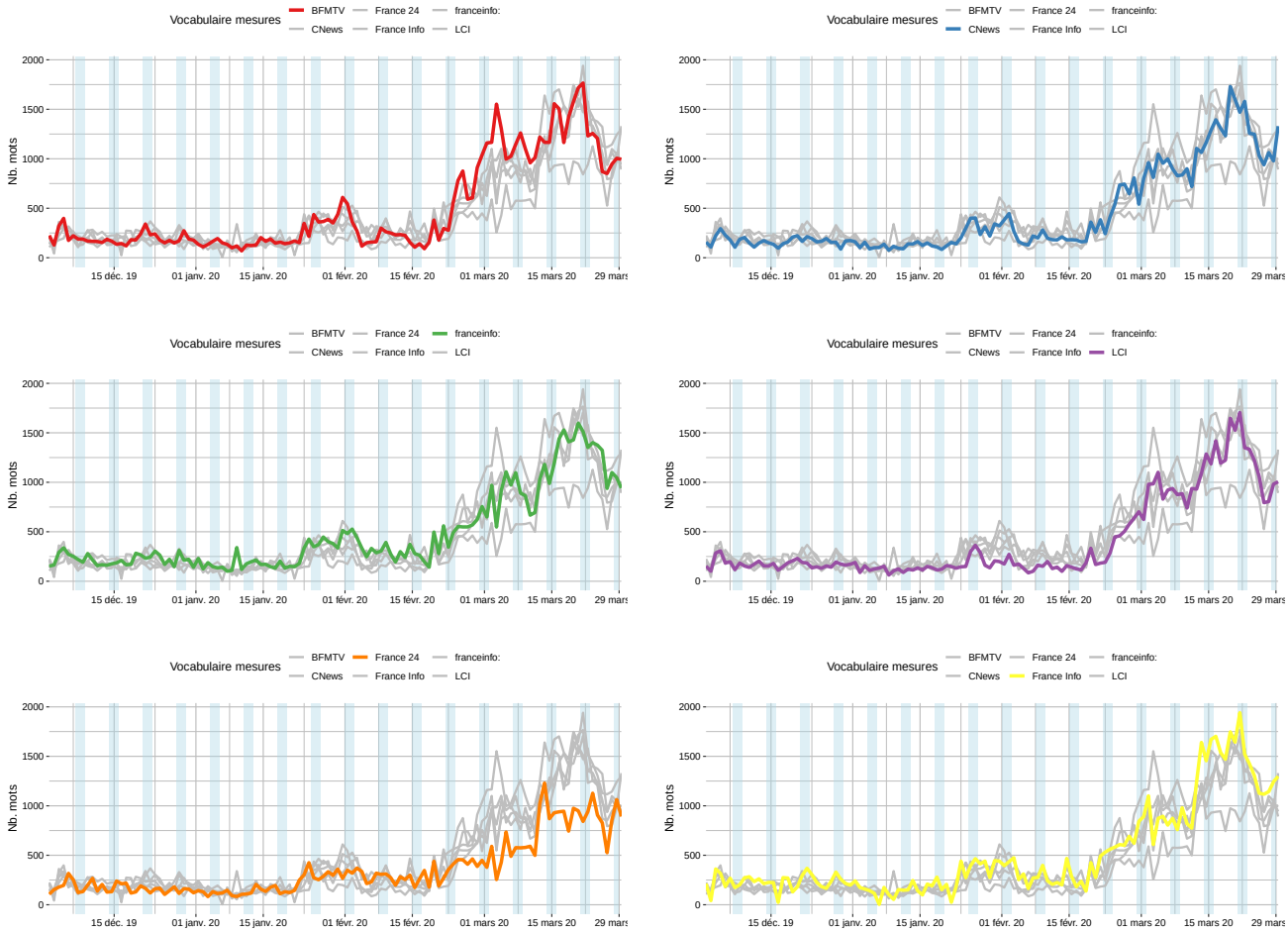


FIGURE 40 – Nombre d'occurrences des mots du groupe *mesures* dans les transcriptions de chaque canal d'information en continu

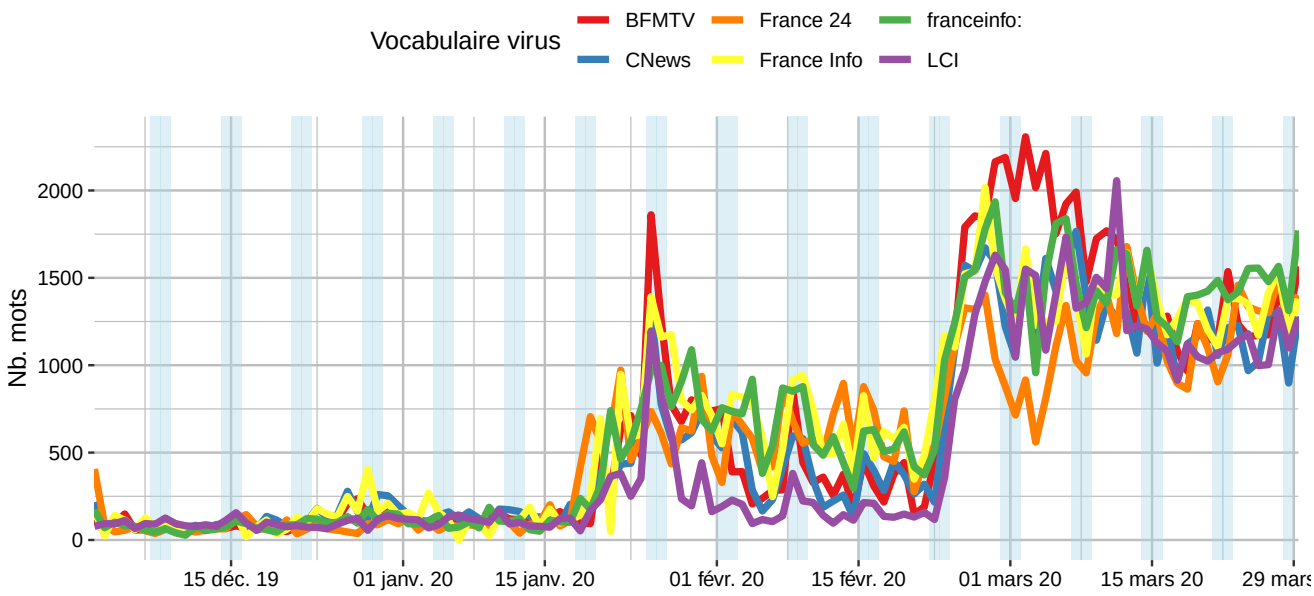


FIGURE 41 – Nombre d'occurrences des mots du groupe *virus* dans les transcriptions de l'ensemble des canaux d'information en continu

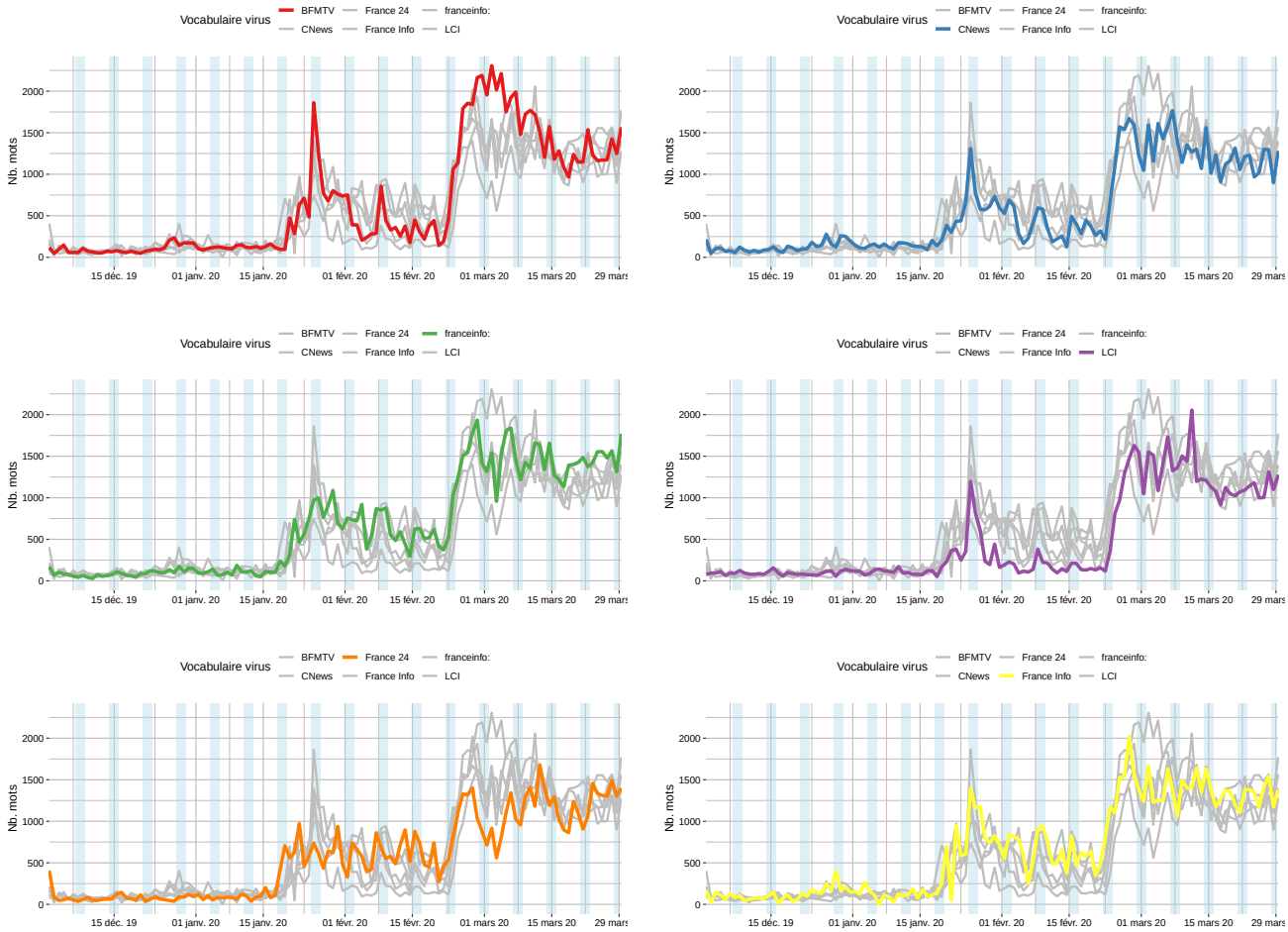


FIGURE 42 – Nombre d'occurrences des mots du groupe *virus* dans les transcriptions de chaque canal d'information en continu

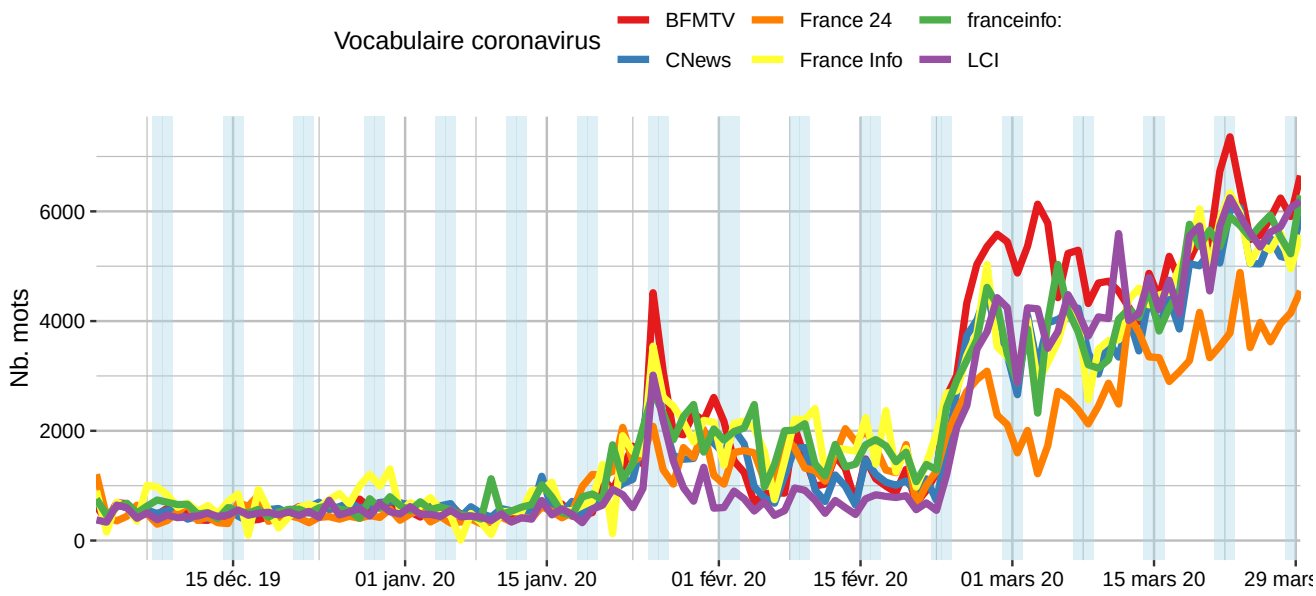


FIGURE 43 – Nombre d'occurrences des mots du groupe *coronavirus* dans les transcriptions de l'ensemble des canaux d'information en continu

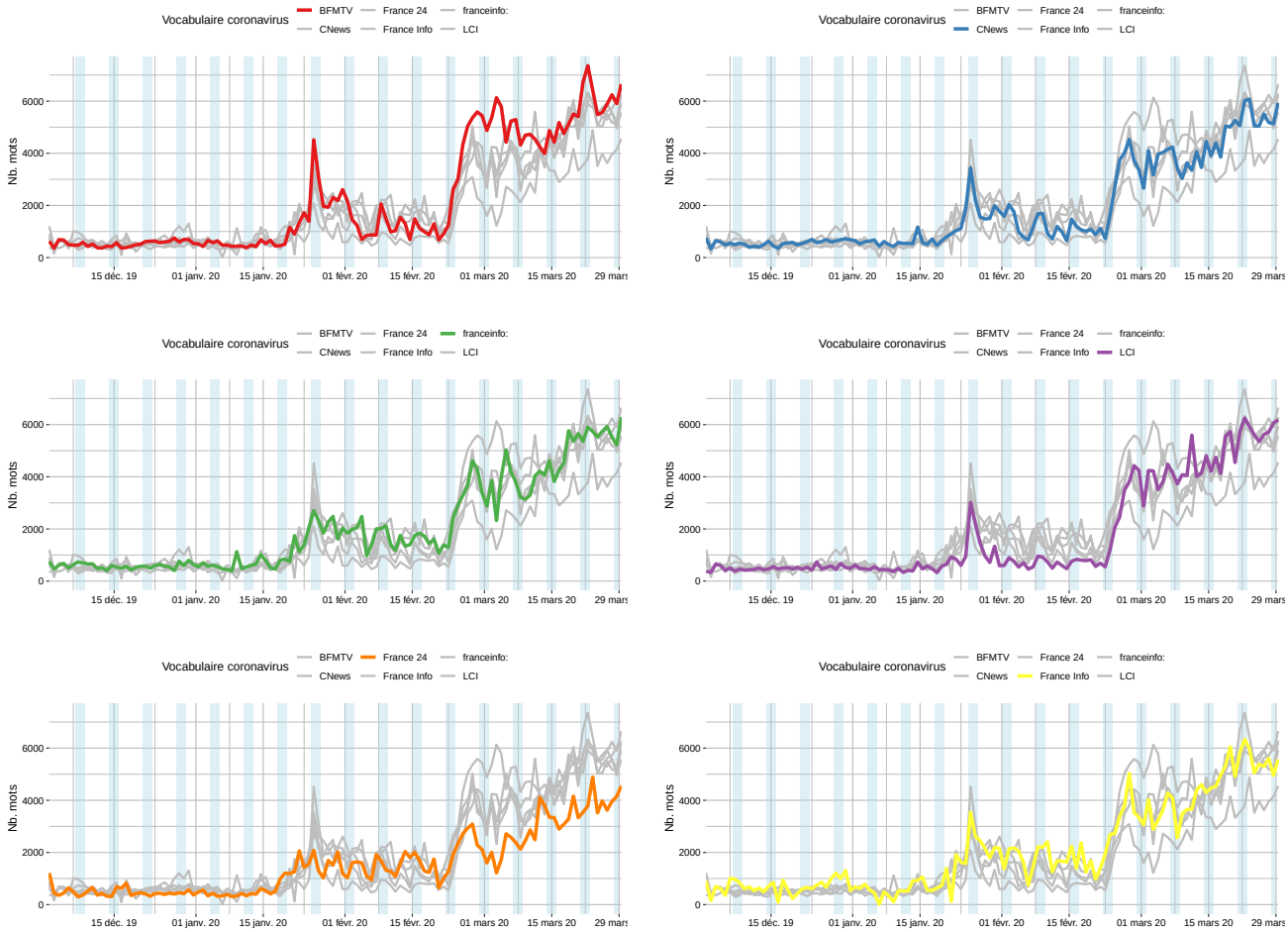


FIGURE 44 – Nombre d'occurrences des mots du groupe *coronavirus* dans les transcriptions de chaque canal d'information en continu

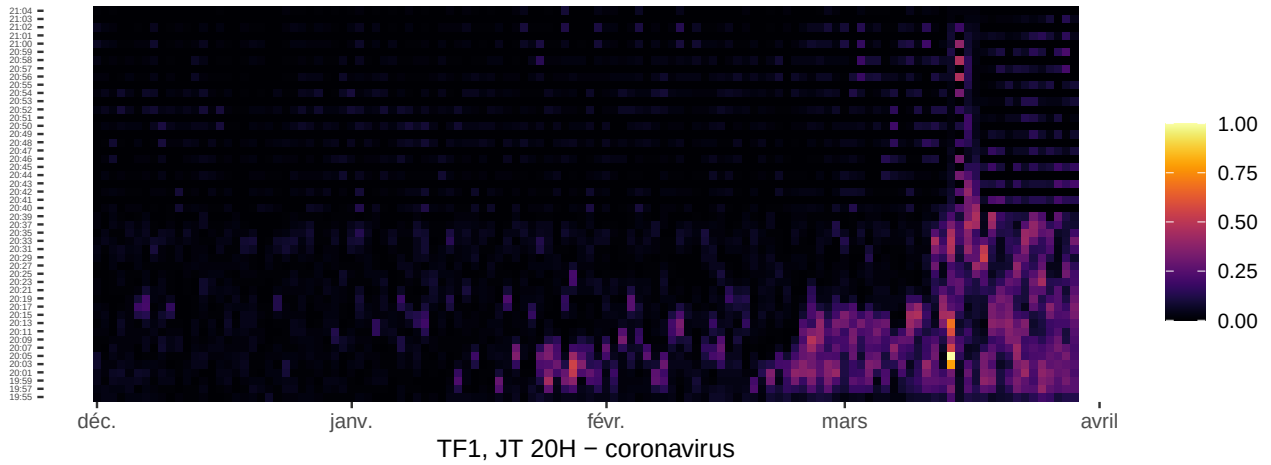


FIGURE 45 – Timeline TF1 coronavirus (normalisation locale)

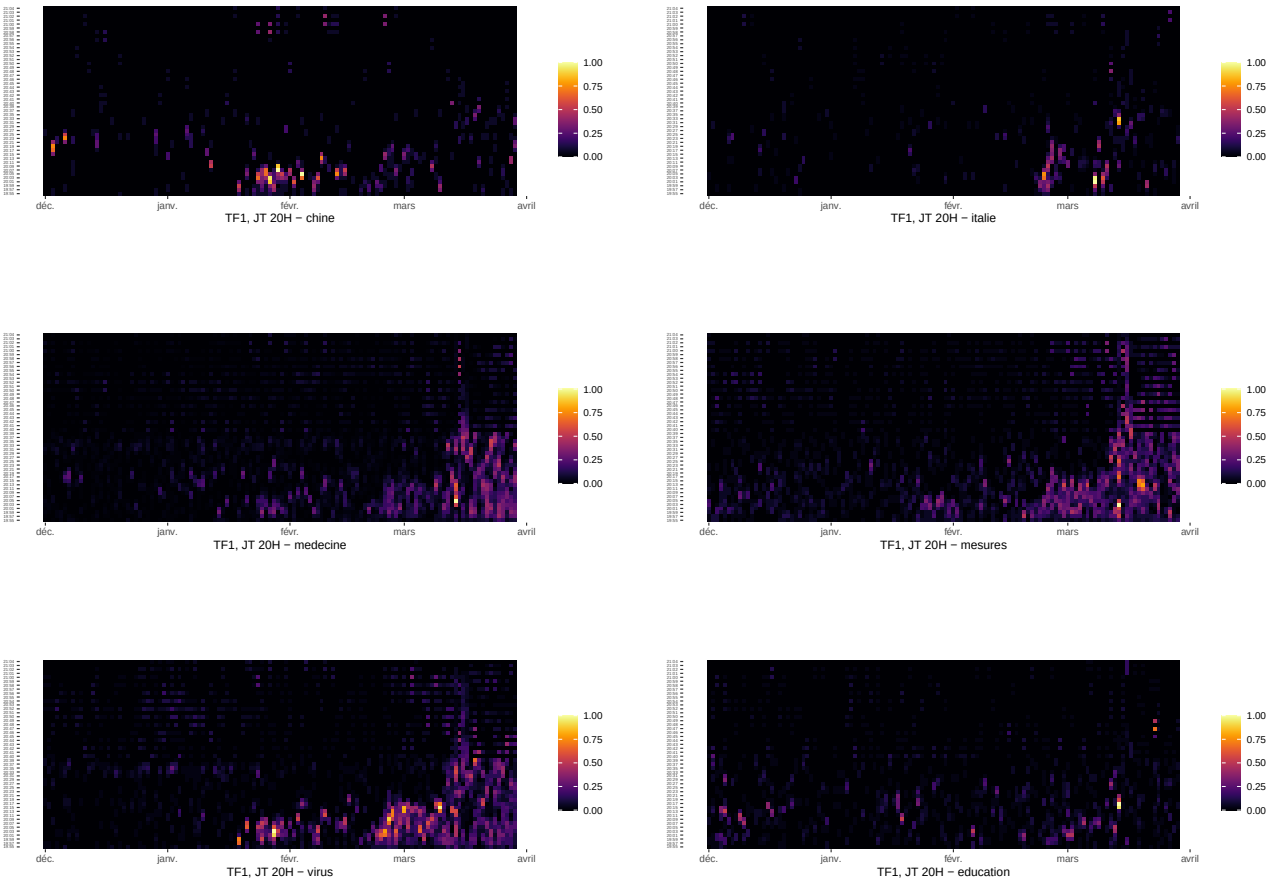


FIGURE 46 – Timeline TF1 autres groupes de vocabulaire (normalisation locale)

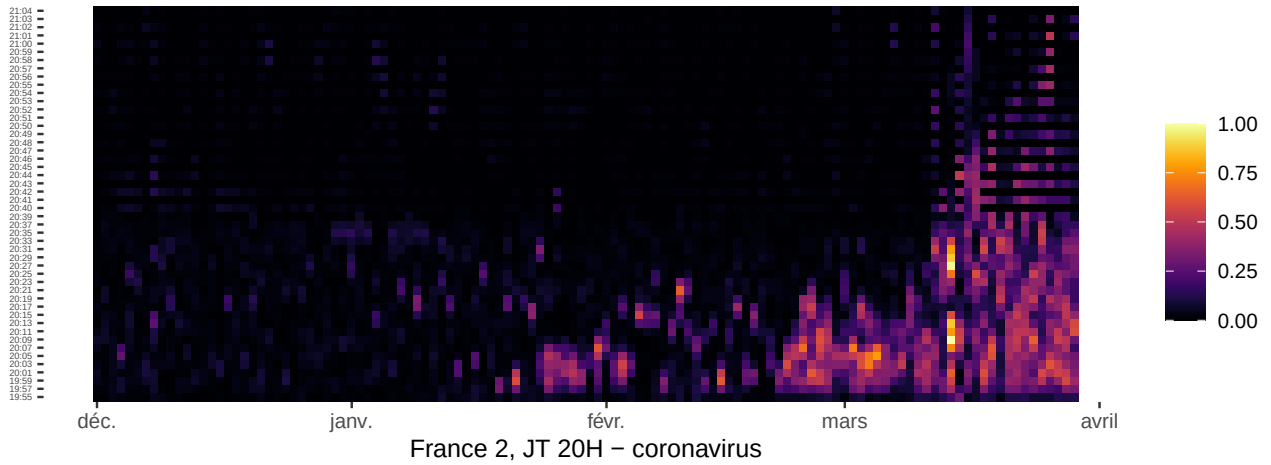


FIGURE 47 – Timeline France 2 coronavirus (normalisation locale)

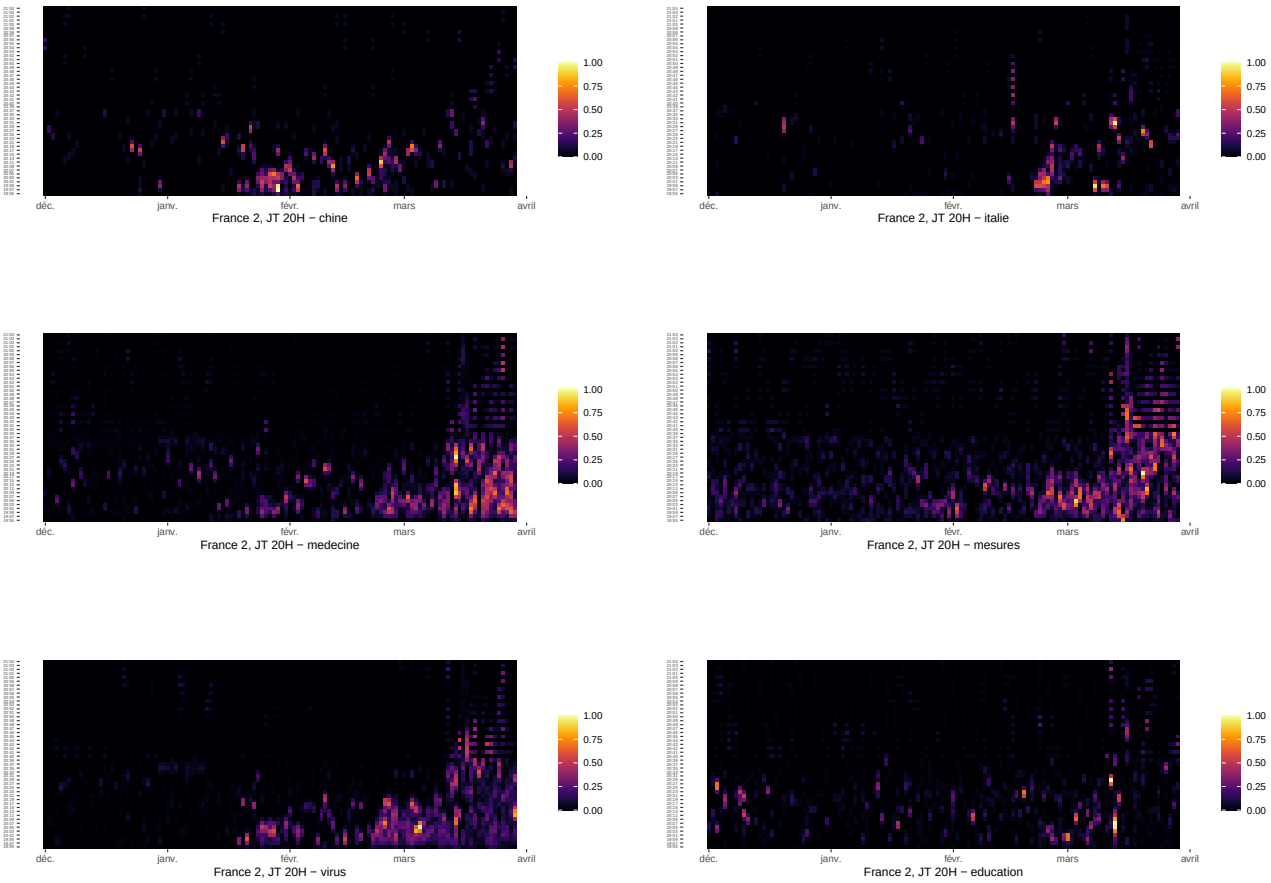


FIGURE 48 – Timeline France 2 autres groupes de vocabulaire (normalisation locale)

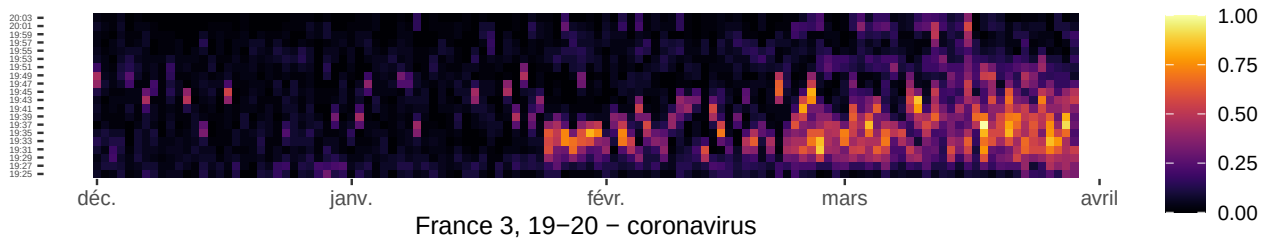


FIGURE 49 – Timeline France 3 coronavirus (normalisation locale)

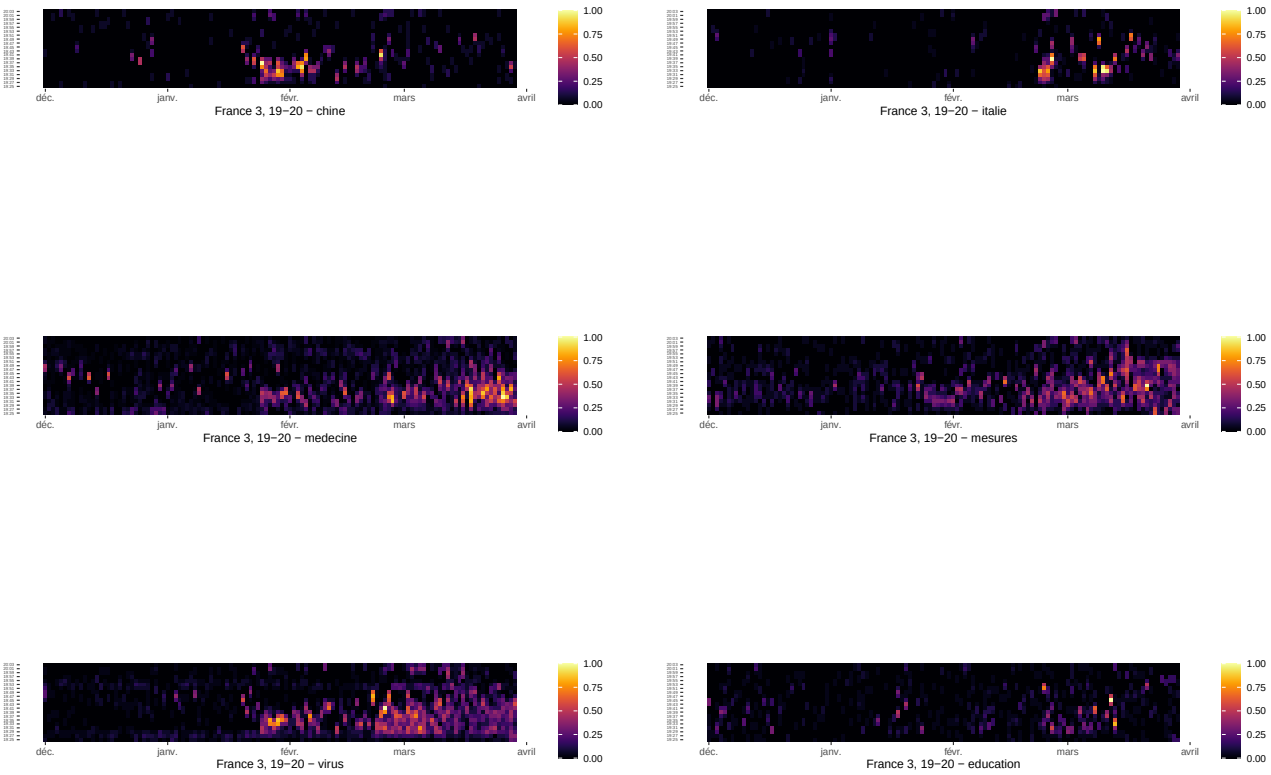


FIGURE 50 – Timeline France 3 autres groupes de vocabulaire (normalisation locale)

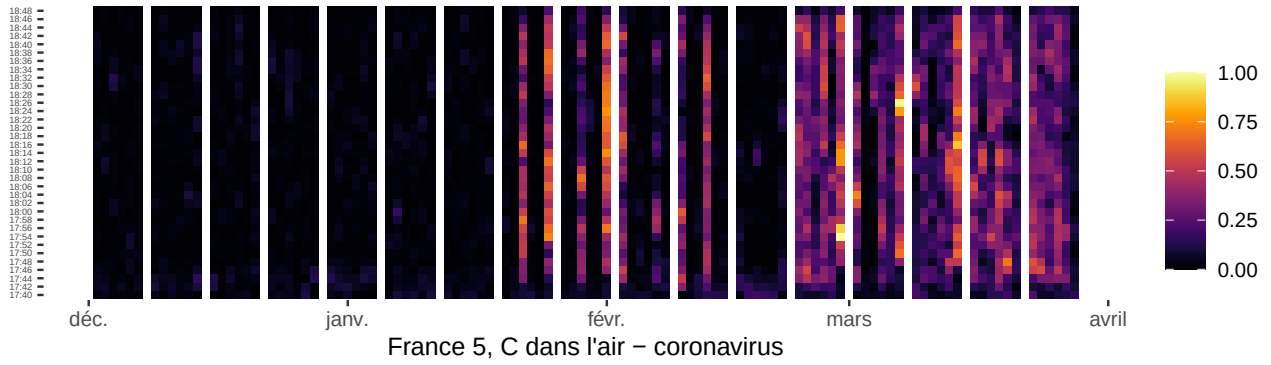


FIGURE 51 – Timeline France 5 coronavirus (normalisation locale)

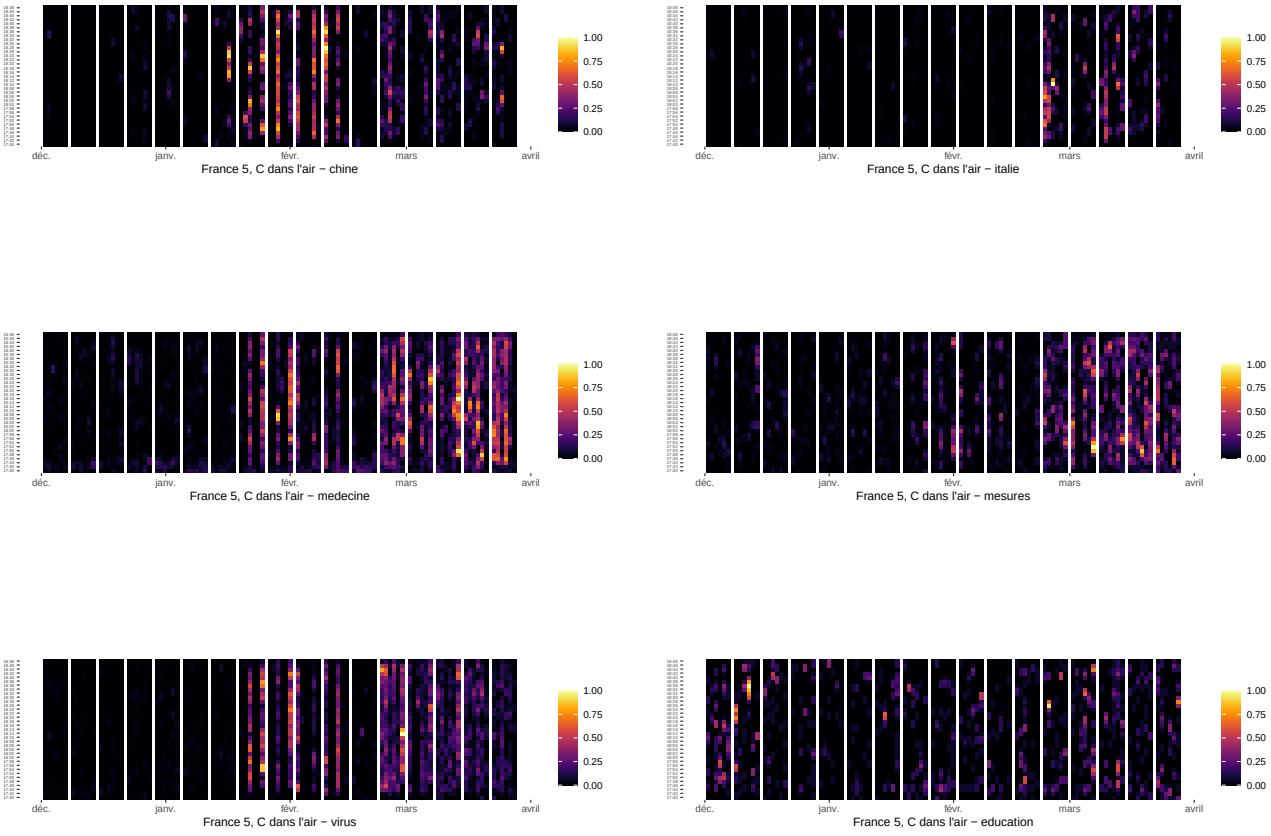


FIGURE 52 – Timeline France 5 autres groupes de vocabulaire (normalisation locale)

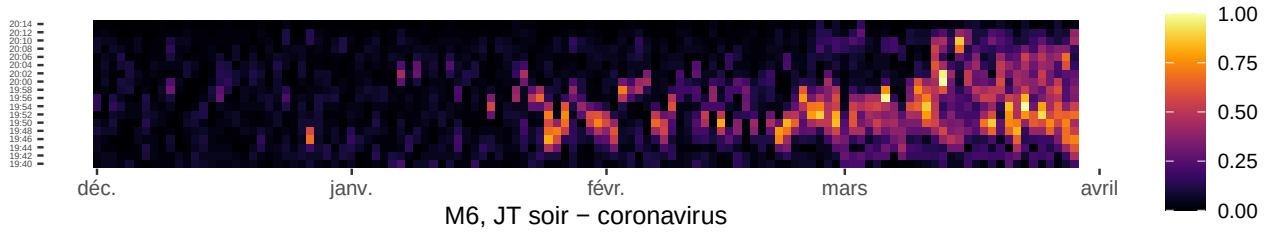


FIGURE 53 – Timeline M6 coronavirus (normalisation locale)

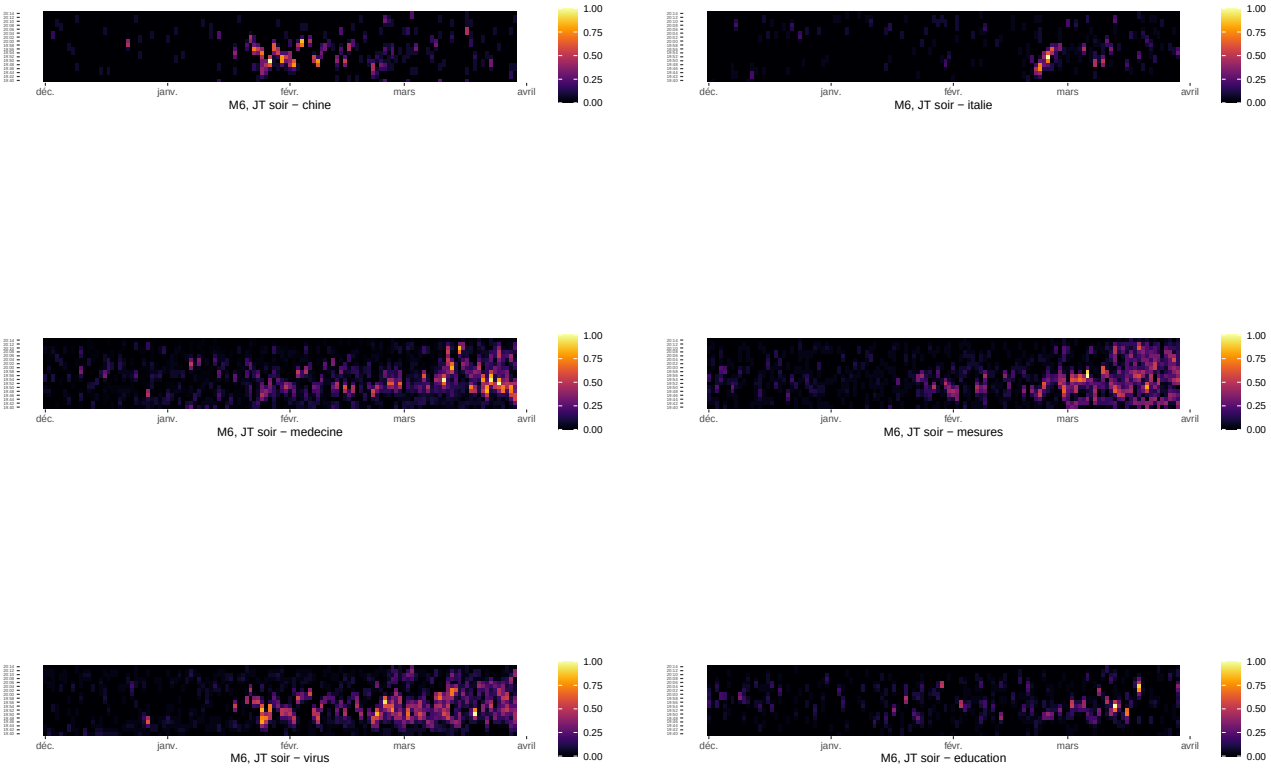


FIGURE 54 – Timeline M6 autres groupes de vocabulaire (normalisation locale)

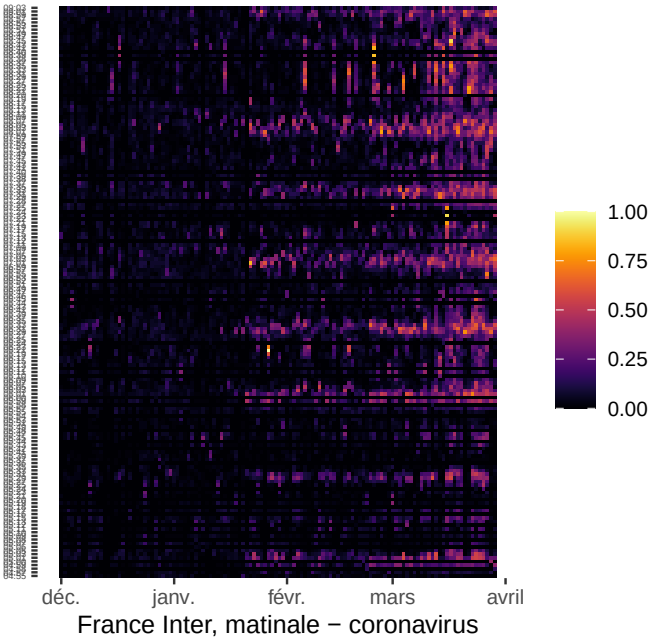


FIGURE 55 – Timeline France Inter coronavirus (normalisation locale)

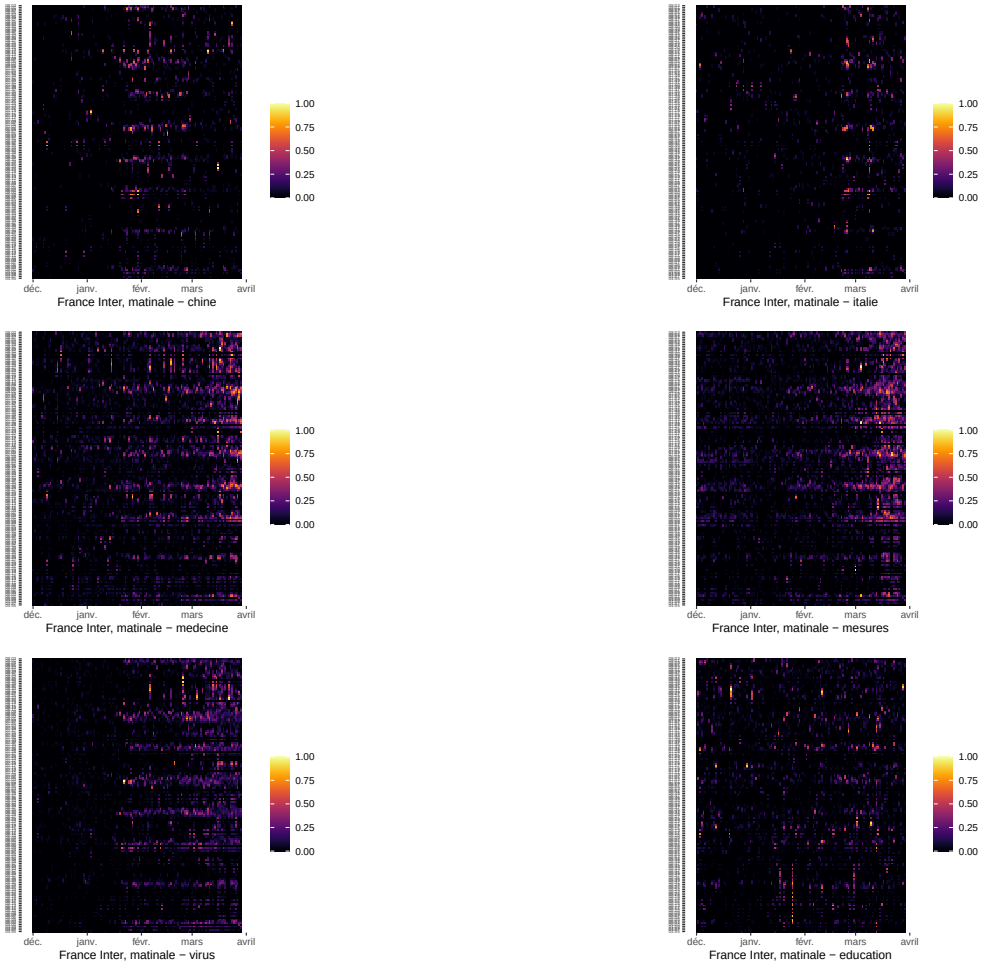


FIGURE 56 – Timeline France Inter autres groupes de vocabulaire (normalisation locale)

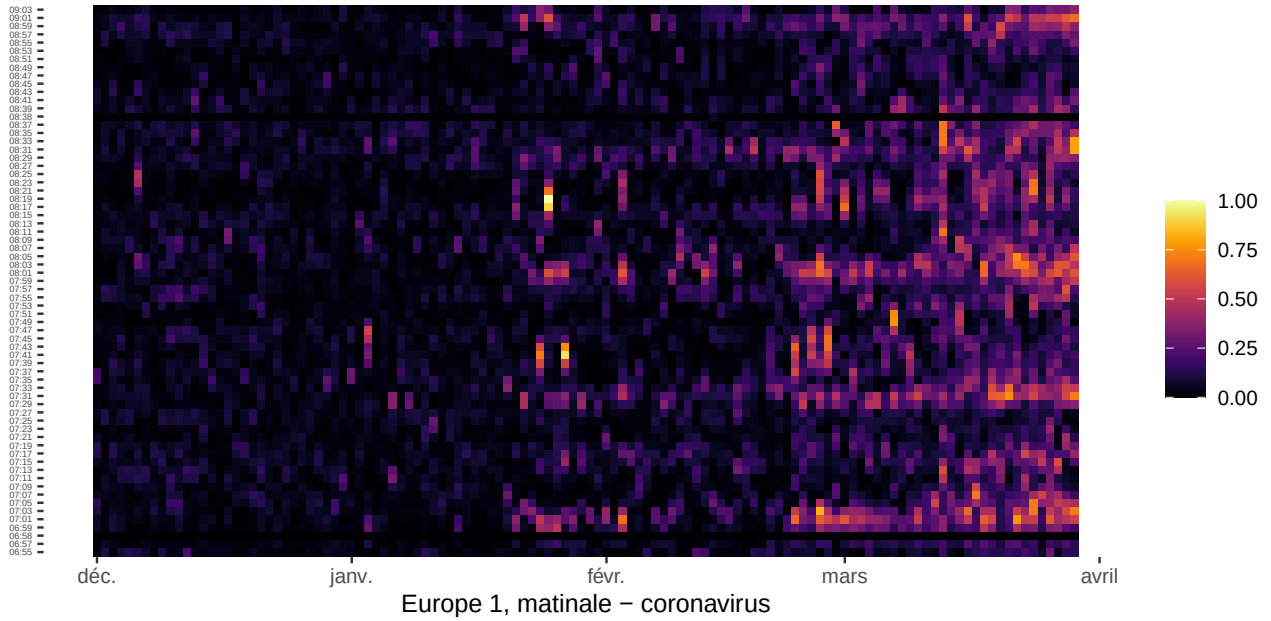


FIGURE 57 – Timeline Europe 1 coronavirus (normalisation locale)

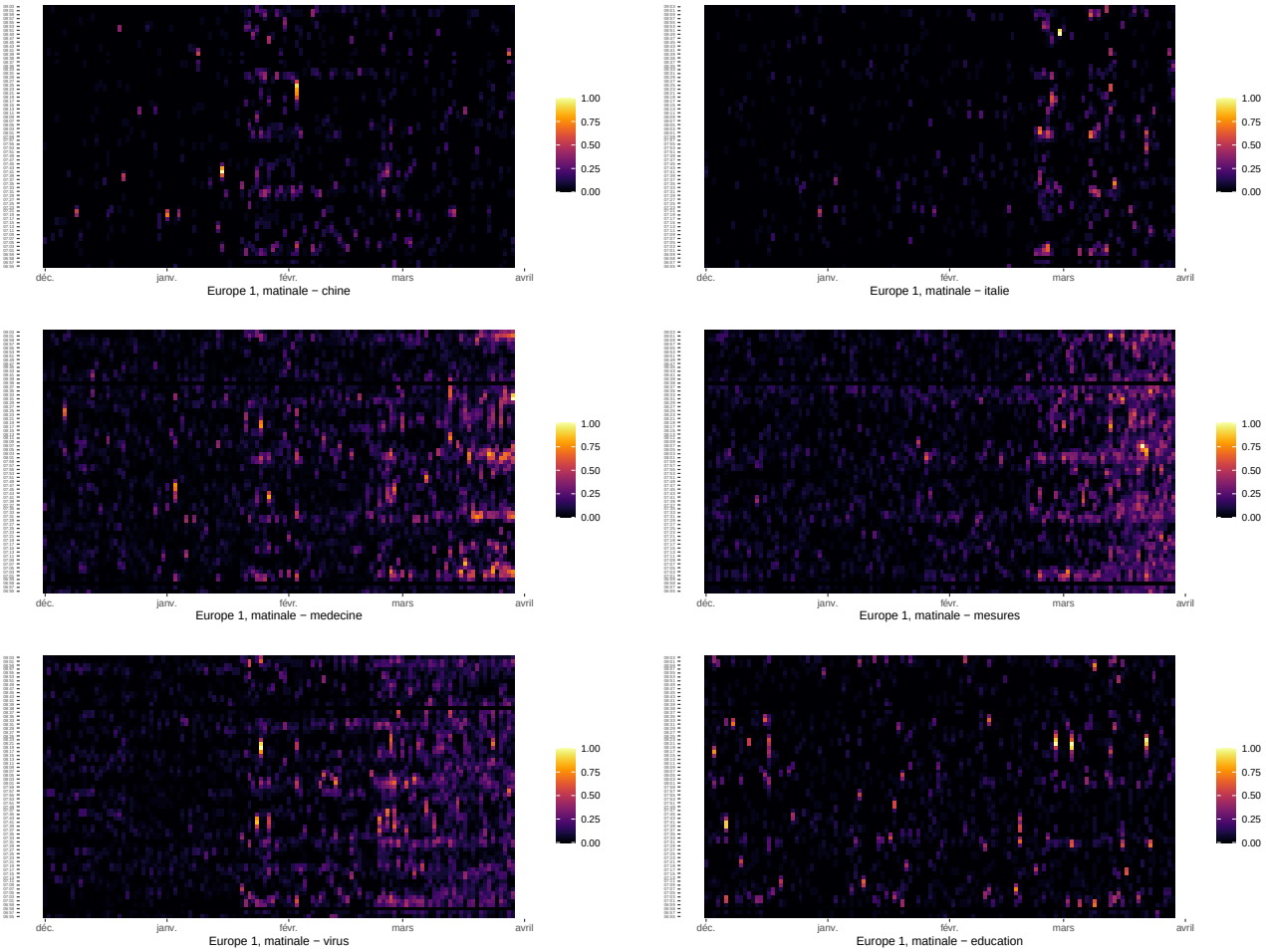


FIGURE 58 – Timeline Europe 1 autres groupes de vocabulaire (normalisation locale)

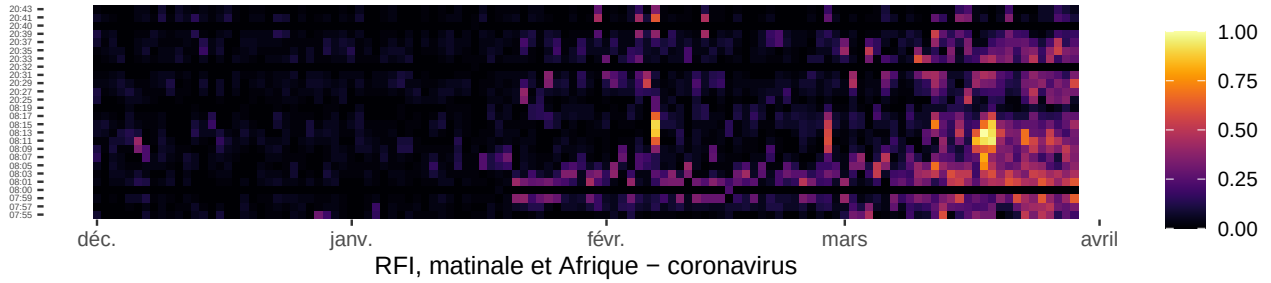


FIGURE 59 – Timeline RFI coronavirus (normalisation locale)

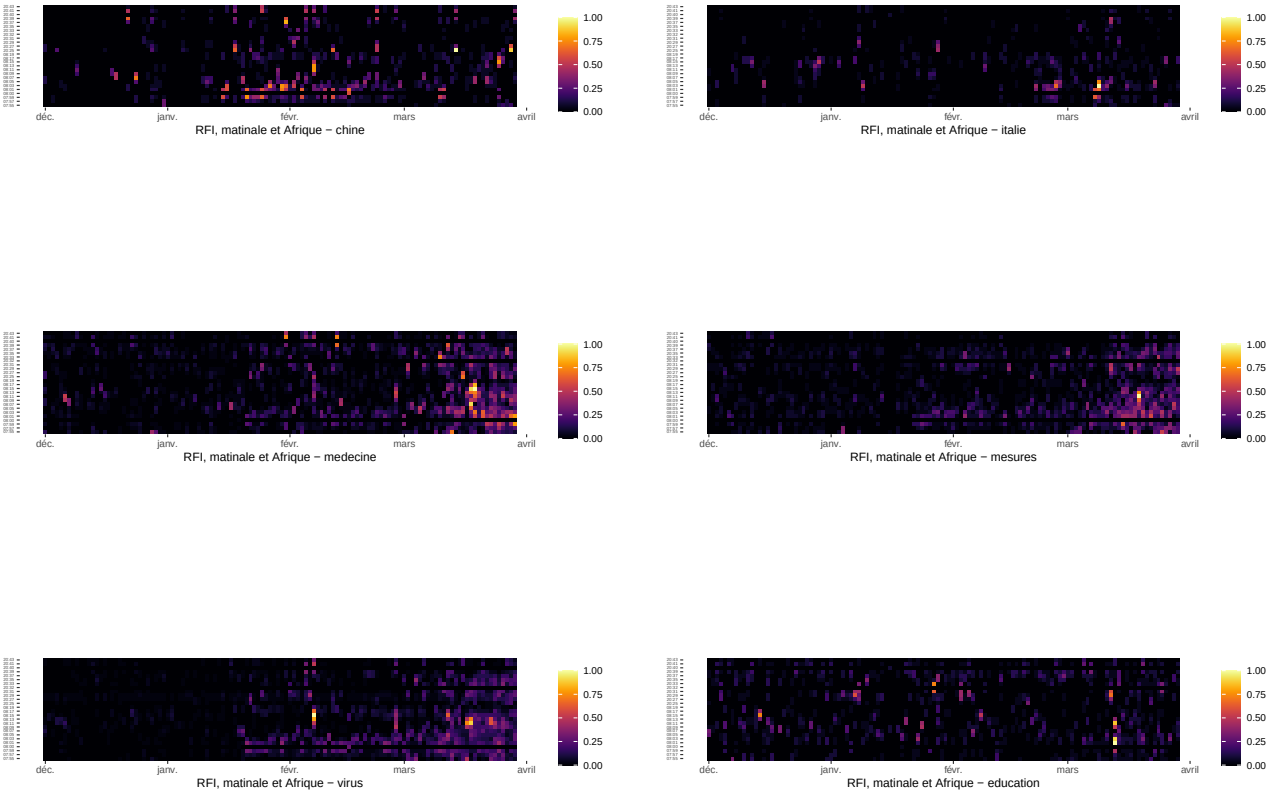


FIGURE 60 – Timeline RFI autres groupes de vocabulaire (normalisation locale)

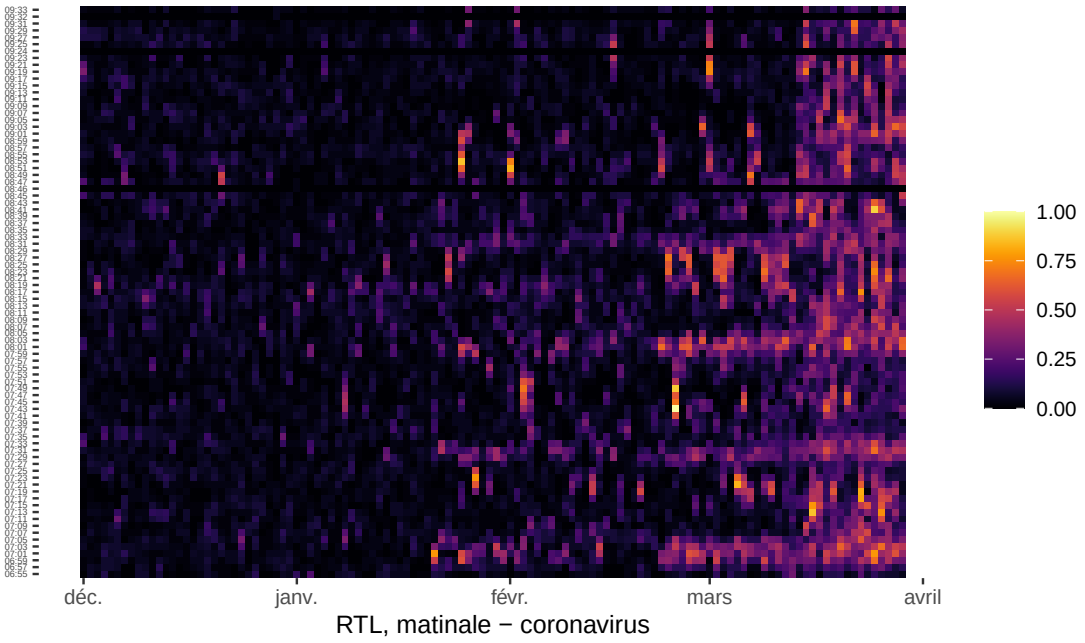


FIGURE 61 – Timeline RTL coronavirus (normalisation locale)

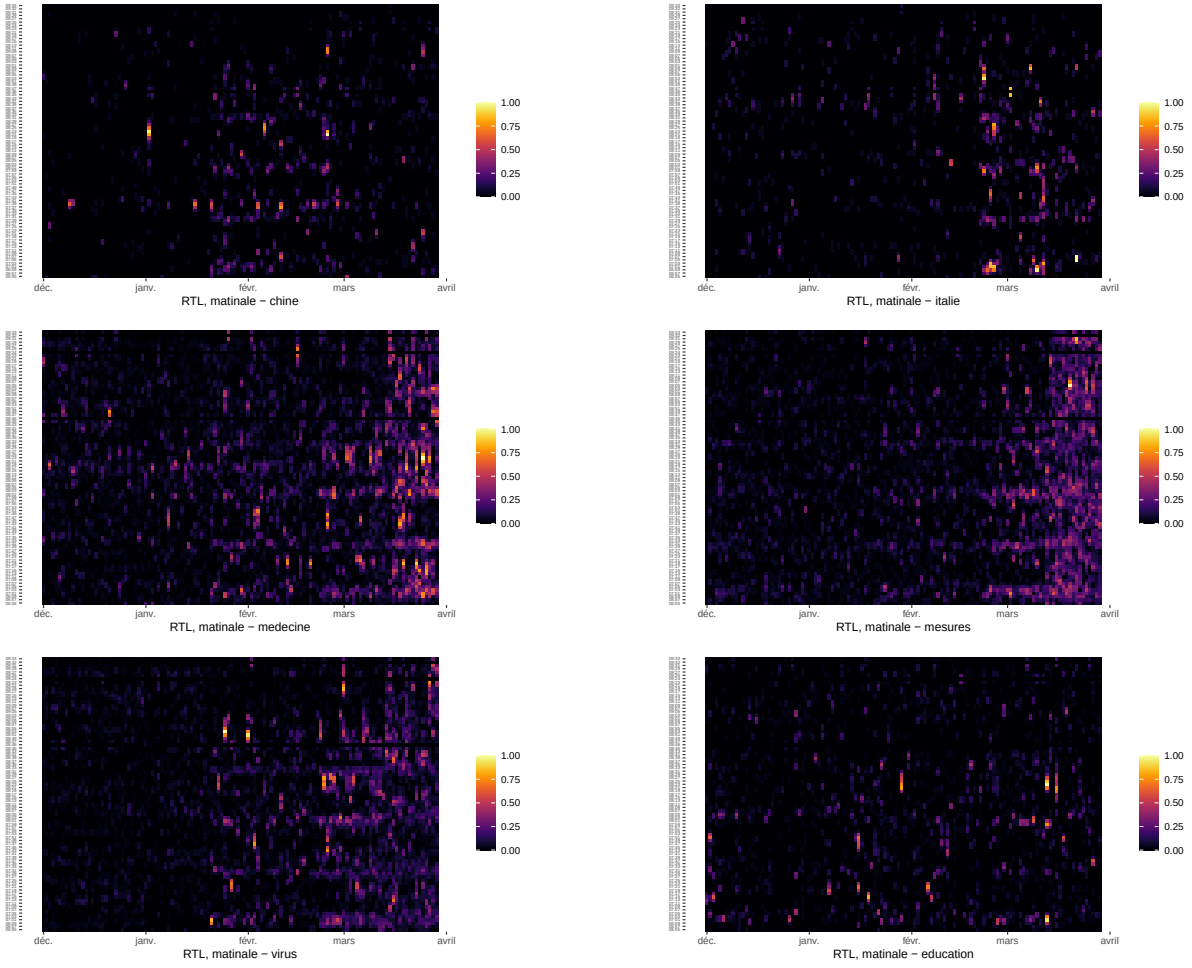


FIGURE 62 – Timeline RTL autres groupes de vocabulaire (normalisation locale)

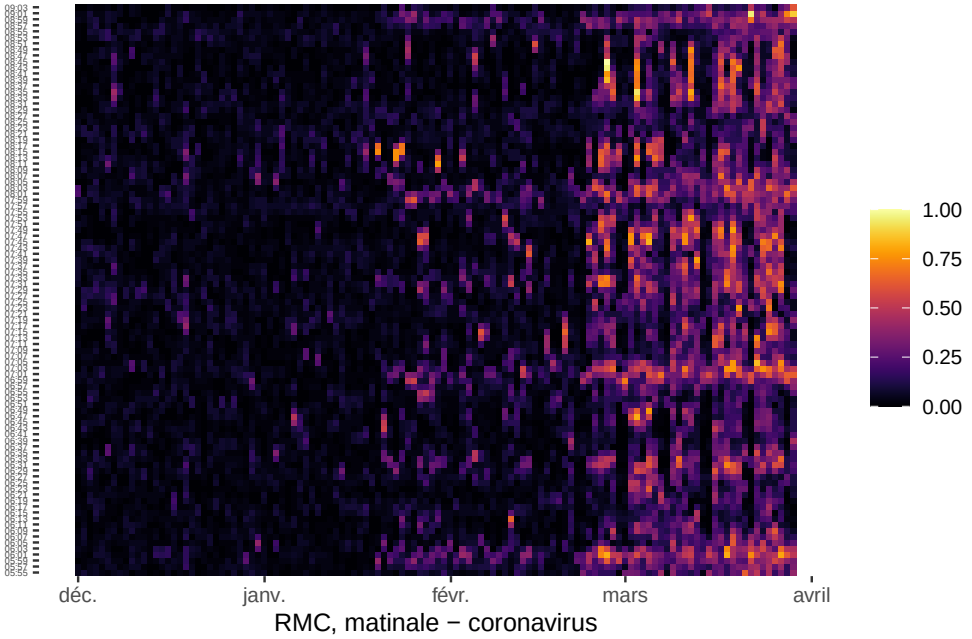


FIGURE 63 – Timeline RMC coronavirus (normalisation locale)

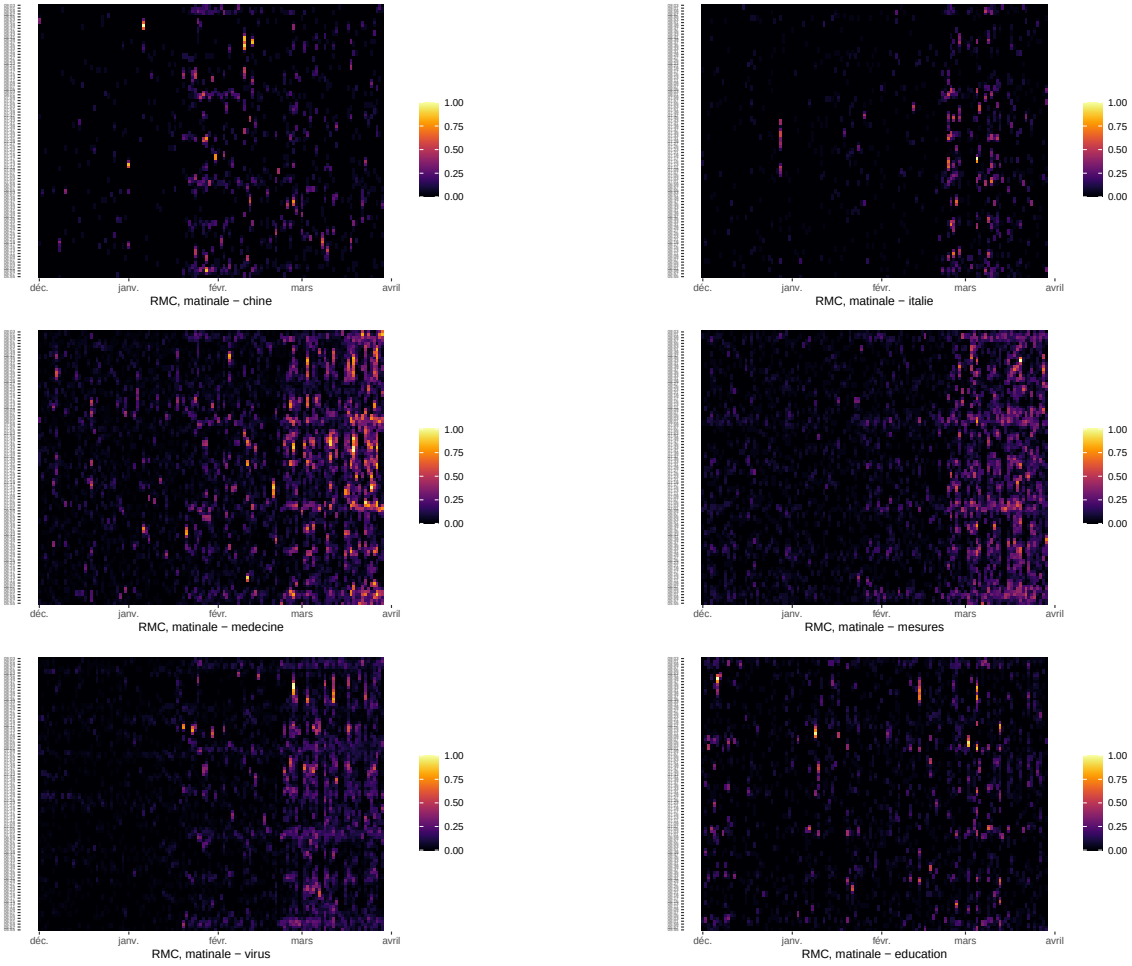


FIGURE 64 – Timeline RMC autres groupes de vocabulaire (normalisation locale)

Table des matières