Cerema

Direction territoriale Ouest

Point d'Appui National Aménagement Numérique des Territoires

Méthode - Base de données «Géolocaux»

Géolocalisation à l'adresse et comptages territoriaux dans le cadre du déploiement des réseaux et services de communications électroniques

Avril 2014

www.ant.developpement-durable.gouv.fr
ant.dvt.DterOuest@cerema.fr

Résumé

Le CETE de l'Ouest [1] et la DATAR [2] ont établi en 2010 un partenariat visant à mettre en place un observatoire national des services de communications électroniques, ayant pour finalité de redistribuer aux collectivités territoriales et services de l'État concernés des informations relatives à la qualité de service de l'internet fixe et à la couverture des populations, entreprises et bâtiments publics du territoire. Ce projet a permis la production d'une base de données géolocalisées sur laquelle se fondent la géolocalisation avancée des données relatives aux niveaux de service de l'internet fixe pour les réseaux filaires et l'évaluation des taux de couverture des foyers et des entreprises de l'observatoire national du Plan France Très Haut Débit.

1/ Production de la base de données des locaux géolocalisés

La méthode se base sur l'hybridation des bases de données de l'IGN (RGE) avec celles de la DGFIP (MAJIC), dans le but d'améliorer la géolocalisation des adresses postales tout en renseignant l'occupation à cette adresse. Les principales étapes du traitement sont les suivantes :

- A partir de la base de données MAJIC, calcul du nombre de locaux d'habitation et professionnels pour chaque adresse postale de chaque parcelle. Les géolocalisants de parcelle MAJIC sont complétés par les géolocalisants de la BDD Parcellaire de l'IGN;
- Pour chacune de ces adresses MAJIC, recherche des points adresses de l'IGN pouvant être appariés de manière à extraire une géolocalisation IGN, pouvant venir compléter le localisant de parcelle;
- Pour chaque adresse MAJIC, en fonction des informations attributaires de la parcelle et des localisants parcelle et adresse, définition d'un point médian, à partir duquel les bâtiments d'accueil les plus probables sont recherchés dans la couche BATI de la BD-TOPO de l'IGN.

La base ainsi produite permet d'obtenir une évaluation fine des répartitions des foyers et des entreprises sur l'ensemble du territoire. Cette base a ensuite fait l'objet d'une analyse pour en déterminer les limites méthodologiques et définir ses possibilités d'utilisation.

2/ Limites méthodologiques et usages

La localisation des foyers et entreprises n'est pas exhaustive (limitée aux locaux fiscalisés) et présente des limites de précisions (quelques dizaines de mètres le plus souvent, jusqu'à 100m dans certains cas). Différents éléments en limitent en effet la précision, en lien avec :

- le contenu et la vocation initiale des bases disponibles (locaux fiscalisés uniquement, bâtiment fiscal sans réalité physique, imprécision en milieu rural, adresses différentes pour une localisation, …) ;
- les traitements réalisés pour faire la jonction entre les deux bases (différences de syntaxe pour une même adresse postale et perturbation des appariements entre adresses MAJIC et IGN, perte de certaines entrées sans correspondance, …) ;
- la technique de recherche du « bâtiment probable » (erreurs cependant minimisées dans les zones peu denses, cible prioritaire du projet) ;

- le cas particulier des grands ensembles (regroupement des locaux sur une seule parcelle de référence dans la base MAJIC, indépendamment de la localisation réelle).

Des contrôles externes et internes ont donc été réalisés et ont notamment établi que la différence entre le nombre de foyers dans la base obtenue et les données de l'INSEE sont de l'ordre de 3,5%, et que les pertes de locaux entre la base obtenue et la base MAJIC initiale étaient inférieures à 0,5%. Ainsi, cette base est pertinente pour toute utilisation conforme à ses objectifs initiaux, à savoir :

- produire des analyses de la couverture numérique qui permettent de disposer d'un outil de diagnostic et de planification, d'aide à la décision et de choix stratégiques ainsi que de suivi des déploiements ;
- alimenter une vision nationale et contribuer à l'évaluation des politiques publiques en matière d'aménagement numérique.

En première approche, on peut donc estimer que la BDD Géolocaux est utile pour toute évaluation visant à réaliser des bilans infra et supra-communaux, pour des données dont la précision de géolocalisation serait de l'ordre de 100m. Par ailleurs, l'utilisation la plus prudente consiste à l'évaluation des taux ou ratio et à éviter les comptages absolus.

Cette BDD expérimentale des locaux fiscalisés géolocalisés offre une vision détaillée de l'occupation du territoire, permettant en outre d'envisager des exploitations dans d'autres cadres que celui de l'aménagement numérique des territoires. Le résultat obtenu présente cependant des limites de précision et de fiabilité, qui si elles ont été estimées compatibles avec les besoins de l'aménagement numérique, pourraient s'avérer incompatibles avec d'autres types de problématiques. La description de la présente méthode peut ainsi servir de base à l'estimation du champ d'application de ce travail.

3/ Conditions d'utilisation

La BDD Géolocaux étant issue de l'hybridation des données du RGE et de MAJIC, son utilisation est par défaut soumise à la somme des obligations liées aux BDD sources. Ainsi, l'usage et la diffusion de la BDD géolocaux doit respecter à la fois les conditions d'utilisation du RGE et les termes de l'engagement CNIL du MEDDE relatif à l'usage des données MAJIC anonymisées.

En particulier, l'usage libre des données est réservé aux missions de service public (par exemple, diagnostics territoriaux nécessaires à la rédaction des documents de planification – SCOT, SDTAN, …). Les missions à caractère industriel et commercial (par exemple toute opération liée au déploiement effectif des réseaux et leur exploitation) ne sont pas autorisées. Pour tout usage ne relevant pas du cadre établi par le RGE et l'engagement CNIL du MEDDE, l'IGN et la DGFIP devront être impérativement consultés.

---

[1] Le CETE de l'Ouest est devenu au 1er janvier 2014 la Direction Territoriale Ouest du Centre d'Etudes et d'Expertise sur les Risques, l'Environnement, la Mobilité et l'Aménagement (Cerema)

[2] la DATAR a intégré en avril 2014 le Commissariat Général à l'Egalité des Territoires (CGET)

Sommaire

1 - CONTEXTE : Aménagement numérique du territoire

Le déploiement de réseaux de communications électroniques à très haut débit jusqu'à l'abonné (FttH) représente un enjeu majeur de développement, tant sur le plan économique que social.

Les opérateurs privés se sont engagés à déployer leurs réseaux optiques dans les zones les plus peuplées, où résident environ 60% de la population française. Équiper le reste du pays, et notamment les territoires les plus ruraux, nécessite l'intervention des pouvoirs publics et particulièrement celle des collectivités territoriales. Faute de cet engagement, certains territoires ne seront pas desservis, générant ainsi une fracture numérique et d'importants déséquilibres économiques mais aussi sociaux. Les collectivités territoriales, soucieuses de l'équilibre économique et social de leurs territoires, ont d'ores et déjà déployé depuis 2004 environ 130 réseaux d'initiative publique (RIP) à haut débit, ce qui représente 3 milliards d'euros d'investissement. Ces déploiements ont permis d'améliorer la couverture en résorbant la plupart des zones blanches (qui représentent désormais moins de 2% du territoire), en subventionnant des modes d'accès alternatifs (satellite) et en favorisant l'apparition d'offres concurrentielles (dégroupage).

Mobilisant un fond d'aide aux collectivités territoriales de 1 milliard d'euros par an, l'État a mis en place dès 2010 un Programme National Très Haut Débit (PNTHD) destiné à soutenir les déploiements. En 2012 le Président de la République a en outre pris l'engagement d'accélérer la cadence de déploiement. La volonté du gouvernement de déployer, sur l'intégralité du territoire, des réseaux optiques destinés à remplacer à terme les actuels réseaux en cuivre, s'appuie ainsi sur la création en décembre 2012 d'une Mission Très Haut Débit, chargée de la mise en œuvre du Plan France Très Haut Débit concrétisant la stratégie de l'État. Pendant les 10 prochaines années, ce Plan mobilisera 20 milliards d'euros, dont 3 milliards en subvention pour soutenir les projets des collectivités territoriales.

Dans ce cadre, la DATAR et le CETE de l'Ouest ont établi en 2010 un partenariat visant à mettre en place un observatoire national des services de communications électroniques, pour produire des supports SIG pour l'aide à la planification des projets de réseaux d'initiative publique (RIP) et permettre le suivi des déploiements par l'État. Ce projet a notamment permis la production d'une base de données géolocalisées, servant de socle à l'observatoire national du Plan France Très haut Débit, sous pilotage de la Mission Très Haut Débit. C'est sur cette base de données que se fondent la géolocalisation avancée des données relatives aux niveaux de service de l'internet fixe pour les réseaux filaires et l'évaluation des taux de couverture des foyers et des entreprises, produites par l'observatoire national du THD.

2 - Besoin de géolocalisation des adresses et des locaux

Les cartes de couverture en services numériques des réseaux filaires (réseau cuivre pour le DSL et réseau fibre pour le FttH) sont produites à partir de points correspondants à des locaux, au droit desquels sont connues les performances à cartographier. Ces dernières[1] concernent :

- les classes de débit (<512Kb/s, entre 512Kb/s et 2 Mb/s, entre 2 et 10 Mb/s, etc.).
- la nature des services disponibles (Internet seul, Téléphonie sur IP, Télévision sur IP).

L'utilisation d'une base de données des locaux géolocalisés doit permettre de déduire des taux de couverture en service numérique de la population et des entreprises.

Or, particulièrement dans le cas du réseau historique cuivre, la géolocalisation des parties terminales des lignes et donc des locaux connectés est inconnue. Deux possibilités permettent alors de proposer une géolocalisation pour le cuivre :

- la géolocalisation à partir d'éléments amonts du réseau et dont la géolocalisation est connue par ailleurs.
- la géolocalisation sur la base de l'adresse postale (cette dernière étant connue au droit des lignes).

2.1 - Géolocalisation imprécise du cuivre par des éléments amonts du réseau

En effet, les différentes lignes terminales, irriguant l'ensemble des locaux, présentent un point de convergence nommé le point de concentration (PC). Il est ainsi possible de choisir de géolocaliser l'ensemble des lignes liées au niveau de leur PC.

Cette méthode s'avère pertinente en site dense, en raison de la faible longueur des tronçons terminaux (longueur de cuivre entre le PC et le local connecté). L'erreur de géolocalisation est alors de quelques dizaines de mètres en moyenne et s'avère compatible avec les exigences du dispositif réglementaire.

En site rural, en revanche, les tronçons terminaux et donc l'imprécision de géolocalisation peuvent atteindre voir dépasser ponctuellement le km.

2.2 - Géolocalisation à l'adresse imprécise en milieu rural

Pour pallier l'imprécision de la méthode précédente, une géolocalisation à l'adresse a été envisagée. Les limites principales de cette méthode sont :

- le taux d'échec d'appariement syntaxique,
- l'imprécision de géolocalisation des BDD adresses existantes en site rural.

[1] précisées dans le décret n° 2009-166 relatif à la publication des services de communications électroniques

Appariement syntaxique :

La géolocalisation à l'adresse est produite par appariement syntaxique entre les adresses des locaux connectés fournies par les opérateurs et les adresses figurant dans la BDD adresses utilisée pour la géolocalisation.

Les taux d'échec pour l'appariement syntaxique se situe classiquement entre 10 et 15% en fonction des départements étudiés. Ce taux à été estimé acceptable dans la mesure ou les cartes surfaciques telecom sont produites à partir d'une très grande quantité de points (autant que de lignes téléphoniques). Les 80 ou 85% restant (représentant 25 millions de points pour la France) couvrent ainsi suffisamment bien le territoire pour permettre la réalisation de cartes compatibles avec les objectifs de l'observatoire.

Limites des BDD adresses actuelles :

Restent les limites liées à la complétude des BDD adresses. Les BDD adresses disponibles à court terme et contenant des informations de géolocalisation et compatible avec les contraintes du projet étaient :

- la BDD Majic (anonymisée et retraitée par le CETE Nord-Picardie[2]),
- la BDD Adresse de l'IGN.

La BDD Majic retraitée de 2009 présentait une géolocalisation des adresses au centroïde de la parcelle. En site dense, cette information est suffisamment précise pour produire les cartes attendues (à 100m près). En revanche, en site rural, les parcelles s'agrandissant, la précision de géolocalisation chute d'autant. Par ailleurs, certaines parcelles ne sont pas géolocalisées.

La BDD adresse de l'IGN présente une géolocalisation des adresses de qualité, sauf pour les plus petites communes (< 3600 hab.) où l'effort de géolocalisation n'a pas encore pu être équivalent. Dans les zones les moins renseignées, l'ensemble des points adresses peut être co-géolocalisé au centre bourg ou sur un lieu-dit voisin.

Or, le champ d'action des réseaux d'initiative publique et donc les besoins d'information pour les collectivités territoriales, s'appliquent particulièrement aux secteurs les moins denses et les plus mal géolocalisés.

2.3 - Erreur de calcul des taux de couverture

Enfin, l'exigence réglementaire, s'appliquant aux opérateurs de communications électroniques et donc indirectement au projet d'observatoire mentionné, porte sur la capacité à calculer des taux de couverture de la population ou des locaux à partir des cartes polygonales services.

Les taux de couverture sont évalués par comptage des locaux partageant le même niveau de performance à l'échelle de la commune la plupart du temps.

Pour ce faire, il faut être en capacité de croiser des cartes vectorielles valides avec une base de données des locaux d'habitation et à usage professionnel, géolocalisés (à 100m près).

[2] Le CETE Nord-Picardie est devenu au 1er janvier 2014 la Direction Territoriale Nord Picardie du Centre d'Etudes et d'Expertise sur les Risques, l'Environnement, la Mobilité et l'Aménagement (Cerema)

Par ailleurs, la non-concordance des sources de géolocalisation peut conduire à des évaluations totalement erronées. Dans le cas où des locaux seraient géolocalisés d'une manière (par exemple sur la base de la couche BATI de la BD Topo de l'IGN) et les adresses d'une autre manière (géolocalisation au PC ou à l'adresse en site rural), les zones cartographiées peuvent ne pas correspondre avec les locaux correspondant. Dans ces cas le taux de couverture ne présente plus aucune fiabilité.

Exemple en zone peu dense:

Zone de couverture évaluée à partir de la géolocalisation des points de concentration du réseau cuivre (■)

Zone de couverture évaluée à partir de la BDD adresse IGN native (●)

Dans ces deux cas, le taux de couverture sera sous-évalué, car la zone habitée estimée (en couleur) ne couvre pas l'ensemble des bâtiments.

3 - Une tentative de réponse : la BDD Géolocaux

En réponse à ces deux limites principales (géolocalisation limitée des adresses pour les zones les plus rurales et absence de BDD de locaux occupés géolocalisés), il a été envisagé de concevoir une BDD utilisable à la fois pour la géolocalisation à l'adresse et pour la connaissance de l'occupation du territoire.

Pouvoir répondre à cette double exigence par l'intermédiaire d'une seule BDD est par ailleurs gage de cohérence dans la mesure ou la géolocalisation de la performance et la géolocalisation des locaux à couvrir sont identiques. (Dans ce cas, le déplacement d'un local ou un défaut de précision de géolocalisation ne modifie pas le taux de couverture résultant).

La BDD recherchée a été obtenue par hybridation des données MAJIC et RGE.

Ce projet a été initié courant 2010 et exploite les livraisons 2011 pour le RGE et 2009 pour MAJIC.

3.1 - Principe de production de la BDD Géolocaux

L'objectif est ainsi de produire une BDD des adresses au droit desquelles sont connus principalement :

- **la géolocalisation**
- **le nombre de locaux d'habitation**
- **le nombre de locaux à usages professionnels**

Les nombres de locaux sont estimés pour chaque adresse de chaque parcelle à partir de la BDD Majic.

La géolocalisation est estimée en plusieurs étapes, à partir des géolocalisants parcelles, des points adresses de l'IGN et de la position des bâtiments extraite de la couche BATI de l'IGN.

La méthode se fonde en effet sur plusieurs sources de géolocalisation de manière à produire une géolocalisation affinée ou palliative. L'association de la géolocalisation à la parcelle et de la géolocalisation à l'adresse permet de pallier l'absence de l'une ou de l'autre ou à produire une géolocalisation hybride. La géolocalisation palliative ou hybride permet ensuite d'optimiser la recherche des bâtiments probables d'accueil afin de rapprocher la géolocalisation de l'adresse du ou des bâtiments correspondants.

Remarque : A titre d'information, le rapprochement ou l'association du point adresse et du point bâtiment correspondant, fait à présent également parti des préoccupations intégrées dans la génération de la BDD adresse du RGE.

Pour les besoins propres à la cartographie du numérique, ce principe à plusieurs objectifs :

- il permet de rapprocher la géolocalisation de la performance du point réel (local) où elle est délivrée.
- il permet d'affiner fortement la géolocalisation de la performance pour les grandes parcelles (où le bâtiment d'accueil peut ainsi être éloigné du centre de la parcelle ou du point adresse quand il existe).
- il permet de découpler les géolocalisations de Majic dans le cas ou plusieurs adresses seraient associées à la même parcelle.

Pour la première version de cette BDD, s'agissant d'une production pour la France entière, un principe d'uniformité de la méthode et des résultats a été recherché. Seules les données disponibles sur l'ensemble du territoire ont ainsi été mobilisées. Pour cette raison et pour cette première version, le PCI vecteur a été écarté (seuls les centroïdes des parcelles sont utilisés).

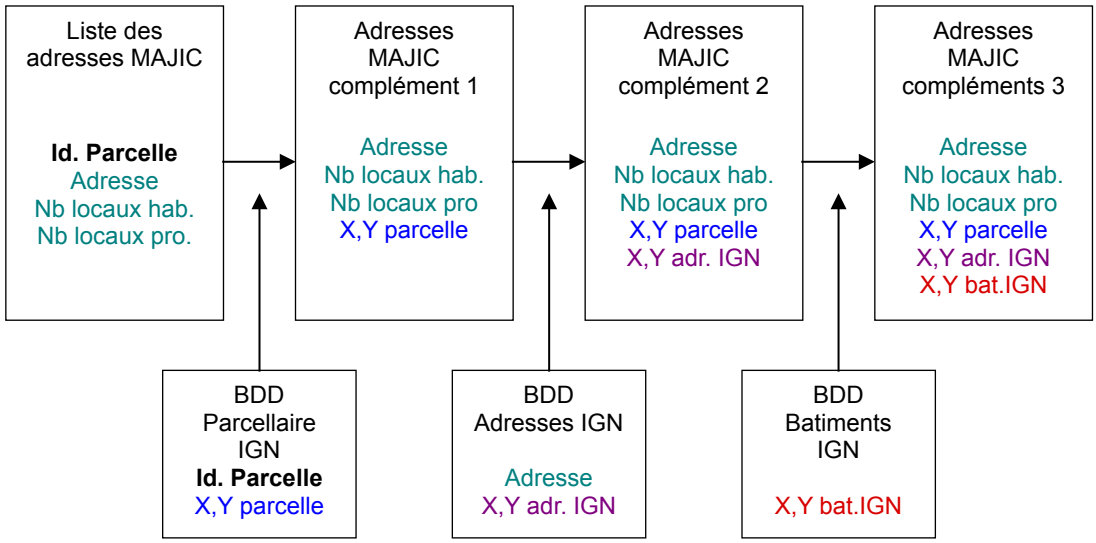
3.2 - Méthode de production de la BDD Géolocaux

Ne sont ici esquissées que les grandes lignes de la méthode. Des éléments plus détaillés pourront être trouvés en annexe.

Les étapes principales du traitement sont :

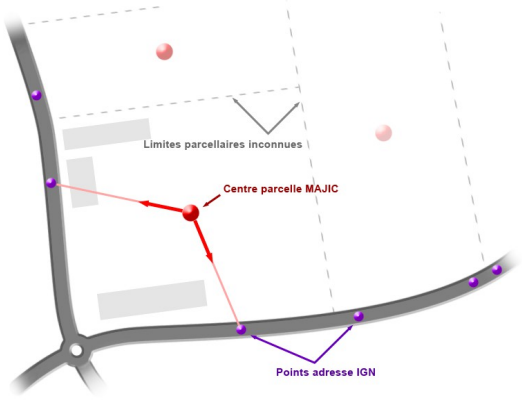
- à partir de la base de données MAJIC, calcul du nombre de locaux d'habitation et professionnels pour chaque adresse postale de chaque parcelle. Les géolocalisants de parcelle MAJIC proviennent de la BDD Parcellaire de l'IGN.
- pour chacune de ces adresses MAJIC, recherche du point adresse de l'IGN pouvant être apparié (par reconnaissance syntaxique) de manière à extraire une géolocalisation adresse IGN, pouvant venir compléter le localisant de parcelle.
- pour chaque adresse MAJIC, en fonction des informations attributaires de la parcelle et des localisants parcelle et adresse, définition d'un point médian, à partir duquel les bâtiments correspondants les plus probables sont recherchés dans la couche BATI de la BD-TOPO de l'IGN.
- à partir d'une distribution des locaux sur les bâtiments probables détectés, calcul des géolocalisations finales (localisation au bâtiment le plus probable, localisation barycentrique des bâtiments probables).

Le schéma ci-dessous représente de façon simplifiée les différentes étapes de la méthode :

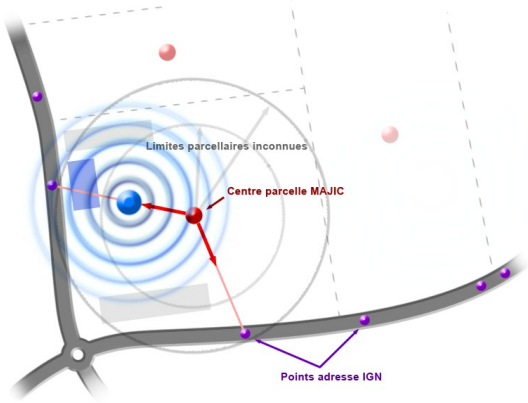


Les différentes étapes de traitement peuvent être illustrées par la série de schémas suivants :

A/ Première série d'étapes conduisant à la génération d'un point médian

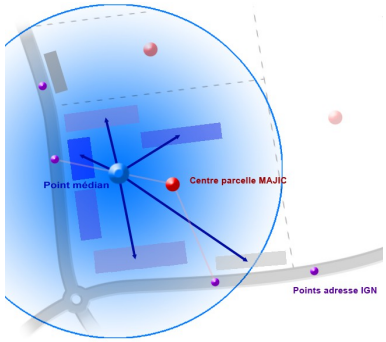


Relier les localisants de parcelle et d'adresse via appariement syntaxique

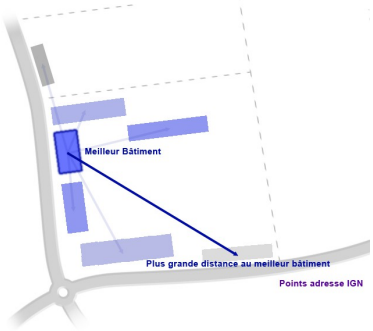


*Création d'un point médian entre localisant de parcelle et d'adresse, probablement à l'intérieur ou proche de la parcelle**

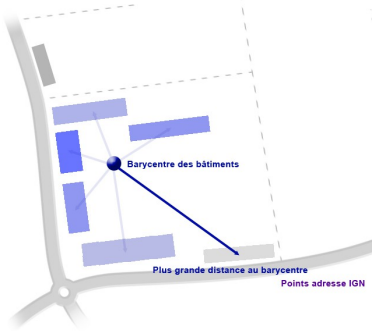
B/ Deuxième série d'étapes conduisant à l'identification du ou des bâtiments probables correspondants



Identification des bâtiments autour du point médian



*Identification du bâtiment recevant le plus de locaux et calcul de la distance maximale aux autres bâtiments identifiés***



Calcul du barycentre et calcul de la distance maximale aux autres bâtiments identifiés

** : La règle de positionnement du point médian est décrite en annexe. Elle vise à tenter de placer le point médian entre le point parcelle et le point adresse tout en restant dans la parcelle supposée (sachant que le contour de la parcelle est, pour ce traitement, inconnu).*

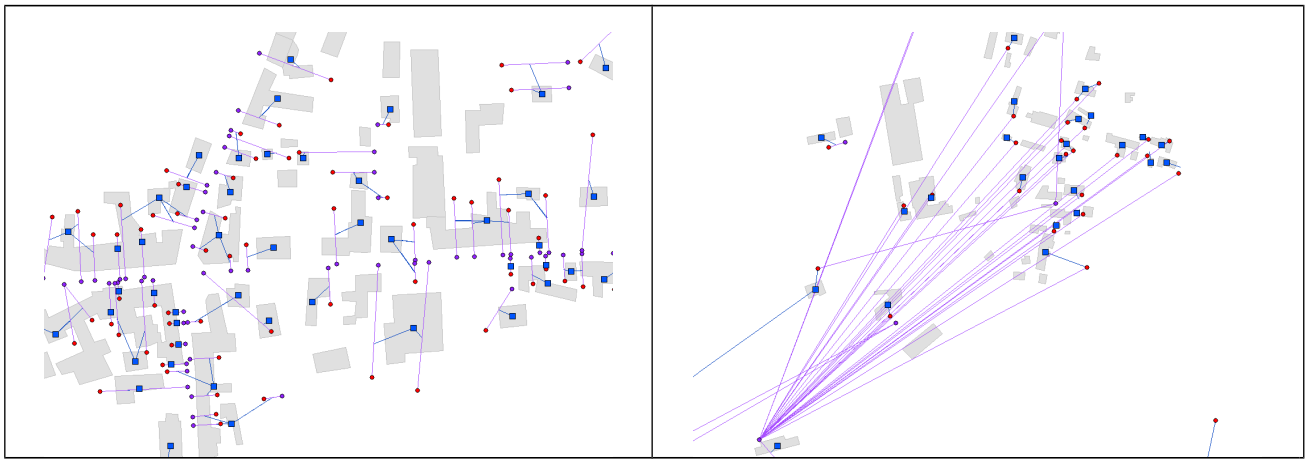
*** : Chaque adresse de chaque parcelle est associée à un nombre de locaux extraits de la BDD Majic. Une règle de distribution entière des locaux au sein des bâtiments probables est décrite en annexe. Le meilleur bâtiment est celui qui reçoit le plus de locaux. Le barycentre est calculé à partir des bâtiments probables en utilisant le nombre de locaux reçus comme coefficient de pondération.*

3.3 - BDD Géolocaux résultante

Globalement les résultats obtenus sont conformes aux attentes.

En particulier, une nette progression de la précision de la géolocalisation des adresses est observée en secteur rural, y compris dans les secteurs ou le PCI vecteur n'est pas disponible (puisque la présente méthode ne l'exploite pas).

Extrait des résultats sur le département 44 :



Exemple en zone dense :
Densité des BDD MAJIC et IGN équivalentes. Le résultat est de même densité mais avec recentrage sur les bâtiments

Exemple en zone peu dense :
La géolocalisation de la BDD Adresse IGN est très nettement complétée par les localisants parcelles et la recherche fructueuse des bâtiments probables

● Point Adresse IGN ● Point Parcelle MAJIC ■ Meilleur Bâtiment identifié

A titre d'information, la production des cartes de services de communications électroniques, exploitant la BDD Géolocaux est également clairement améliorée par rapport aux méthodes initiales.

Exemple en zone peu dense :

Zone habitée évaluée à partir de la géolocalisation des points de concentration du réseau cuivre

Zone habitée évaluée à partir de la BDD adresse IGN native

Zone habitée évaluée à partir de la BDD Géolocaux

Dans le dernier cas, l'évaluation de la zone habitée, sur laquelle se fonde ensuite le calcul des taux de couverture, est mieux répartie sur le territoire et qualifie mieux les masses bâties réelles.

La structure de la BDD Géolocaux (métadonnées) est la suivante :

Nom du champ	type	largeur	nom du champ complet
NomCommune	caractère	28	Nom de la commune
CodeINSEE	caractère	5	Code INSEE de la commune
Adr_Numero	entier		Numéro de l'adresse
Adr_BisTer	caractère	3	Indice de répétition le cas échéant
Adr_Origine	caractère	70	Libéllé Voie ou lieu-dit
NbFoyer	entier		Nombre de logement à cette adresse
NbEntreprise	entier		Nombre de locaux commerciaux ou professionnel à cette adresse
MAJIC_X	réel		coordonnée du localisant parcelle MAJIC complété par la BD_PARCELLAIRE IGN
MAJIC_Y	réel		coordonnée du localisant parcelle MAJIC complété par la BD_PARCELLAIRE IGN
IGNAdress_X	réel		coordonnée du point correspondant de la BDADRESSE® IGN
IGNAdress_Y	réel		coordonnée du point correspondant de la BDADRESSE® IGN
Median_X	réel		Coordonnée du point médian construit à partir du géolocalisant parcelle MAJIC et du point adresse IGN
Median_Y	réel		Coordonnée du point médian construit à partir du géolocalisant parcelle MAJIC et du point adresse IGN
BaryBat_X	réel		Coordonnée du barycentre les bâtiments probables correspondant à cette adresse
BaryBat_Y	réel		Coordonnée du barycentre les bâtiments probables correspondant à cette adresse
BestBat_X	réel		Coordonnée du meilleur bâtiment probable à cette adresse
BestBat_Y	réel		Coordonnée du meilleur bâtiment probable à cette adresse
Adress_Alea	caractère	1	Qualité d'appariement syntaxique entre l'adresse MAJIC et l'adresse IGN *
Adress_XYRedondance	entier		Redondance géographique du point adresse IGN (Nombre de points adresse IGN co-géolocalisés)
Bat_NbAttrib	entier		Nombre de bâtiment probables correspondant à cette adresse
BaryBat_Rayon	entier		Distance maximale entre le barycentre et le plus éloigné des bâtiments probables (en mètres)
BestBat_Rayon	entier		Distance maximale entre le meilleur bâtiment et le plus éloigné des bâtiments probables (en mètres)

* *A = Très bon*
B = Bon
C = Moyen
D = Mauvais ou impossible

4 - *Limites de la BDD Géolocaux*

4.1 - Décomptes limités aux locaux fiscalisés

La présente démarche n'exploitant que les données du RGE et de MAJIC, seuls les locaux fiscalisés sont comptabilisés. Ainsi, **les locaux et équipements publics ne sont pas pris en compte**.

Pour l'usage visé, dans le champ du numérique, cette limite n'est pas préjudiciable. En effet, l'objectif est dans le cas présent de calculer des taux de couverture en service internet grand public. Or, la majorité des équipements publics sont dotés de connexions dédiées et ne souscrivent pas à des offres grands publics. Ainsi, les taux de couverture calculés à partir des seuls locaux fiscalisés restent pertinents.

La BDD pourrait néanmoins évoluer en tentant de prendre en compte les équipements publics pouvant être identifiés à partir des données du RGE. En effet, les points d'activités et d'intérêt (PAI) pourraient être sélectionnés en fonction de leur nature et intégrés à la BDD Géolocaux.

4.2 - Perte modérée d'adresses et de locaux

La méthode utilisée se fonde sur la liste des adresses extraite de MAJIC. Pour chacune de ses adresses, un traitement est réalisé afin de tenter d'affiner la géolocalisation.

Dans la mesure où nous avons besoin de géolocaliser des locaux, les adresses IGN qui ne peuvent être appariées avec une adresse MAJIC ne sont pas conservées (puisque leur occupation est inconnue).

Par ailleurs, la BDD Géolocaux ne retient que les adresses ayant pu être géolocalisées. Les adresses MAJIC non géolocalisées (c'est-à-dire attachées à une parcelle non-géolocalisée et ne pouvant pas être appariées avec une adresse IGN dont la géolocalisation est estimée exploitable) sont écartées.

Dans ces rares cas, les locaux associés sont également exclus de la BDD. Cela représente environ 0,3% des locaux répertoriés dans MAJIC.

4.3 - Localisation fiscale des locaux

Les locaux attachés à chaque adresse de chaque parcelle de MAJIC répondent à une logique d'attribution hybride, géographique et fiscale. Il arrive ainsi que certains locaux ne se trouvant pas physiquement sur une parcelle soient localisés par exemple au droit de la parcelle de leur propriétaire.

Il ne semble pas possible de détecter ces glissements d'attribution de manière à corriger la localisation de ces locaux en les replaçant dans la bonne parcelle.

4.4 - Confusion avec les parcelles voisines

Pour cette première production de la BDD Géolocaux, le parti pris d'uniformité nationale nous a conduit à ne pas utiliser les parcelles vectorisées. Ainsi, au droit de chaque adresse, les bâtiments probables sont détectés dans un rayon fonction de la surface fiscale de la parcelle. Sans connaissance des contours de la parcelle, particulièrement lorsque des parcelles très allongées sont accolées, il arrive que soit pris en compte les bâtiments des parcelles voisines.

Dans la BDD Géolocaux, la prise en compte des géolocalisations barycentriques à tendance à tempérer (flouter) ce défaut. En revanche la prise en compte de la géolocalisation « au meilleur bâtiment » peut correspondre à un bâtiment de la parcelle voisine. Ce cas de figure se présente essentiellement en site dense. Dans ce cas, l'erreur de géolocalisation reste inférieure au 100m annoncés.

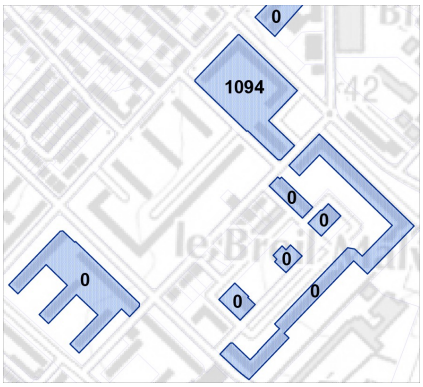
Une amélioration évidente de la BDD pourrait provenir de l'exploitation du cadastre vectorisé. Dans ce cas seraient seulement pris en compte les bâtiments se trouvant à l'intérieur du contour réel de la parcelle.

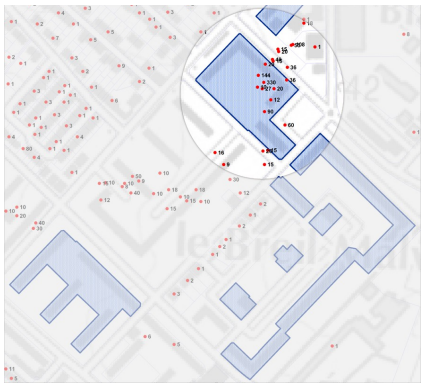
4.5 - Grands ensembles

La comptabilisation de locaux peut être faussée par le contenu des tables MAJIC. Dans la table pnb10_parcelle de MAJIC, les locaux ne sont pas toujours quantifiés en détail sur chacune des parcelles. En effet, dans le cas des grands ensemble et des co-propriétés, les locaux sont agrégés et mentionnés sur la parcelle de référence, au détriment des parcelles qui lui sont liées.

Exemple sur le secteur du Breil Malville (Dept 44) correspondant à un ensemble HLM :

IDPAR	…	**IDPARREF**	IDSECREF	IDCOM	IDCOMTXT	…	NLOCAPPT	NLOCHABIT	NLOCCOM	…
44109000LZ0090	…			44109	Nantes	…	**1094**	**1094**	**2**	…
44109000LZ0094	…	**44109000LZ0090**	44109000LZ	44109	Nantes	…	**0**	**0**	**0**	…
44109000LZ0096	…	**44109000LZ0090**	44109000LZ	44109	Nantes	…	**0**	**0**	**1**	…
44109000LZ0097	…	**44109000LZ0090**	44109000LZ	44109	Nantes	…	**0**	**0**	**0**	…
44109000LZ0103	…	**44109000LZ0090**	44109000LZ	44109	Nantes	…	**0**	**0**	**0**	…
44109000LZ0105	…	**44109000LZ0090**	44109000LZ	44109	Nantes	…	**0**	**0**	**0**	…
44109000LZ0106	…	**44109000LZ0090**	44109000LZ	44109	Nantes	…	**0**	**0**	**0**	…
44109000LZ0113	…	**44109000LZ0090**	44109000LZ	44109	Nantes	…	**0**	**0**	**0**	…
44109000LZ0121	…	**44109000LZ0090**	44109000LZ	44109	Nantes	…	**0**	**0**	**0**	…





On constate que dans ce cas, l'ensemble des locaux est attribué dans MAJIC à une seule parcelle de référence, au détriment des parcelles liées. Ce lien est identifiable via l'attribut IDPARREF.

Malgré les mécanismes de compensation mis en place dans le processus actuel et permettant de distribuer les locaux MAJIC dans les bâtiments de la couche BATI, l'essentiel des locaux attribués reste concentré à proximité de la parcelle de référence.

Une analyse préliminaire des champs IDPAR et IDPARREF permettrait ainsi de détecter ce phénomène, de manière à redistribuer les locaux de la seule parcelle de référence sur l'ensemble des parcelles liées.

La distribution des locaux sur les points de la BDD Géolocaux ne pourra qu'en être améliorée.

5 - Conditions d'utilisation

5.1 - Pour quels usages ?

La pertinence de cette base de données pour le calcul des taux de couverture des foyers et des entreprises en service Internet fixe grand public est assurée, les points où la performance est connue étant géolocalisés par l'intermédiaire de la même BDD. Ainsi, quelle que soit la précision des points, la coïncidence est garantie.

Ceci peut donner lieu à généralisation : la BDD des locaux géolocalisés peut ainsi être estimée pertinente pour toute évaluation exploitant des données géolocalisées à l'adresse par l'intermédiaire de cette même base.

Lorsqu'une carte de performance, dont la couverture est à évaluer, est géolocalisée par une tierce méthode, il est alors nécessaire de connaître la précision absolue de géolocalisation. Cette dernière est en général de l'ordre de quelques dizaines de mètres, voire 100 m. Une telle précision est suffisante pour de très nombreuses évaluations et dans de nombreux champs disciplinaires.

Néanmoins, la BDD des locaux géolocalisés n'est pas exhaustive. En effet, en cas d'échec de géolocalisation, les locaux liés sont écartés. Par ailleurs, ne sont pris en compte que les locaux fiscalisés.

Si l'objectif est l'identification individuelle de foyers et d'entreprises concernées par une quelconque évaluation, la méthode présentée doit être exploitée avec la prudence la plus grande, puisque la géolocalisation individuelle de chaque foyer et de chaque entreprise n'est pas garantie.

Ainsi, en première approche, on peut estimer que la BDD Géolocaux est utile pour toute évaluation visant à réaliser des bilans infra et supra-communaux, pour des données dont la précision de géolocalisation serait de l'ordre de 100m.

Par ailleurs, l'utilisation la plus prudente consiste à l'évaluation des taux ou ratio et à éviter les comptages absolus. *Par exemple, le nombre de locaux par commune n'est pas totalement fiable, en revanche le pourcentage communal représenté par les locaux contenus dans un périmètre infra-communal est relativement fiable.*

5.2 - Conditions d'utilisation RGE et CNIL

La BDD Géolocaux étant issue de l'hybridation des données du RGE et de MAJIC, son utilisation est par défaut soumise à la somme des obligations liées aux BDD sources.

Ainsi, ***l'usage et la diffusion de la BDD géolocaux doit respecter à la fois les conditions d'utilisation du RGE et les termes de l'engagement CNIL du MEDDE relatif à l'usage des données MAJIC anonymisées.***

Annexes

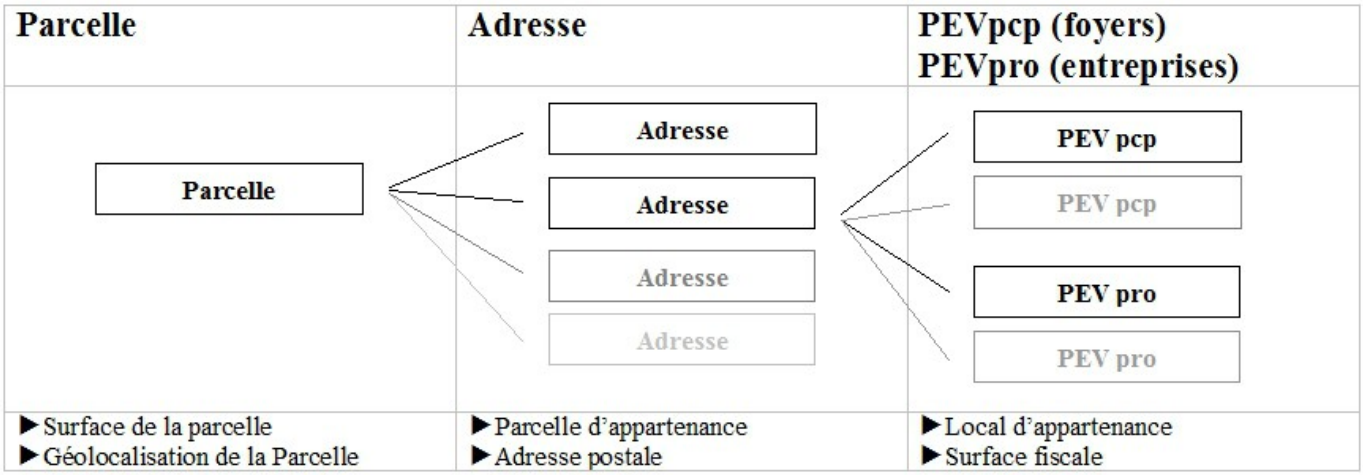
Annexe 1 : Méthodologie détaillée de production de la base

A/ Extraction de la BDD MAJIC

Seules sont exploitées les tables et champs des tables MAJIC permettant de fournir une évaluation du nombre de locaux hab. et nombre de locaux pro. pour chaque adresse de chaque parcelle. Des informations de surface sont extraites au vol de manière à faciliter les étapes ultérieures.

De manière synthétique, l'organisation des données MAJIC est la suivante :

- une parcelle comprend des locaux, ces derniers sont dénommés : « parties d'évaluation » (PEV) correspondant à des locaux d'habitation ou a vocation professionnelle,
- chacune de ces couches est dotée d'informations attributaires exploitables :



- un ensemble de locaux situés à une adresse au sein d'une parcelle pourra correspondre statistiquement, mais sans systématisme, à un bâtiment construit identifiable.

Seules les parties d'évaluation correspondant à un niveau de taxation principale sont conservées concernant les locaux d'habitations, alors que tous les niveaux de taxation sont conservés pour les locaux professionnels. Ces choix sont relatifs aux exigences propres à l'aménagement numérique. En effet, dans ce cadre, seuls sont recherchés les locaux susceptibles d'accueillir une connexion Internet fixe grand public. Sont ainsi notamment écartés les bâtiments publics, ces derniers bénéficiant de connexions dédiées. Le cas échéant, la présente BDD pourrait être complétée par les données extraites des points d'activités et d'intérêt (PAI) de la BD TOPO.

B/ Complément IGN des géolocalisants de parcelle

Des informations concernant la géolocalisation des parcelles cadastrales sont également disponibles à partir de la base de données Parcellaire de l'IGN.

En raison des mises à jour asynchrones de la DGFIP et de l'IGN, la liste des géolocalisants de parcelles ne sont pas équivalentes. Ainsi, des géolocalisants de parcelles de l'IGN peuvent être ajoutés à ceux de MAJIC, de manière à partir d'une base de données la plus complète possible.

C/ Liste d'adresses MAJIC par agrégation

L'objectif premier étant de concevoir une base de données affinées des adresses à des fins de géolocalisation, le niveau maximal d'agrégation efficace est celui de l'adresse MAJIC (et non la parcelle, puisque cette dernière peut présenter plusieurs adresses).

La structure de la base de données obtenue pour cette première étape est donc la suivante :

Adresses Postales MAJIC :

- ► Nombre de locaux d'habitation (nombre de PEV pcp)
- ► Nombre de locaux professionnels (nombre de PEV pro)
- ► Surface fiscale de la parcelle d'appartenance
- ► Surface fiscale bâtie agrégée (somme des surfaces fiscales des PEV)
- ► Adresse postale
- ► Géolocalisant de la parcelle d'appartenance

D/ Liens entre les adresses MAJIC et IGN pour ajout d'un localisant

L'affinement des géolocalisations ne peut être obtenu que par multiplication des sources de données. L'objectif est ici de compléter le géolocalisant de la parcelle MAJIC par celui des points adresse de l'IGN.

L'appariement entre adresse MAJIC et adresse IGN ne peut se faire que par reconnaissance syntaxique de l'adresse postale.

Adresse MAJIC		**Adresse IGN**
►Adresse postale ►Géolocalisant parcelle	⟷	►Adresse postale ►Géolocalisant IGN

En raison des différences de syntaxe et des hétérogénéités de saisie, cet appariement n'est pas trivial et a nécessité le développement d'outils de traduction/comparaison permettant de calculer la « distance syntaxique » rapidement.

Ainsi, pour chaque adresse MAJIC, la méthode consiste à trouver l'adresse IGN présentant la distance syntaxique la plus faible. Si cette dernière est inférieure à un seuil alors l'appariement est possible et le géolocalisant parcelle se voit complété par le géolocalisant du point adresse IGN apparié.

Nota 1 : *Cette phase est particulièrement consommatrice de temps. Pour cela, un algorithme rapide de calcul des distances syntaxiques a été développé.*

Nota 2 : *Dans le cas où le numéro de voie exacte n'est pas trouvé au sein de la BDD adresse de l'IGN, un géolocalisant est interpolé entre les numéros de voirie immédiatement inférieur et supérieur.*

Nota 3 : L'appariement syntaxique intègre une phase de traduction conforme des adresses dont les phases principales sont :

- *suppression de tous les accents*
- *mise en majuscule de toutes les lettres*
- *suppression des articles supposés non discriminants*
- *mise en conformité des abréviations courantes*
- *mise en conformité et reconnaissance des types de voie et lieux (place, rue, etc.).*

A l'issue de ce traitement, la base de données résultante présente la structure suivante :

Adresses Postales MAJIC :

- ► Nombre de locaux hab.
- ► Nombre de locaux pro.
- ► Surface fiscale de la parcelle d'appartenance ($S_{Parcelle}$)
- ► Surface fiscale bâtie agrégée ($S_{Bat\ MAJIC}$)
- ► Adresse postale conforme
- ► Géolocalisant de la parcelle d'appartenance ($X,Y_{Parcelle\ MAJIC}$)
- ► Géolocalisant du point adresse IGN apparié ($X,Y_{Adresse\ IGN}$)
- ► Distance syntaxique IGN-MAJIC
- ► Niveau de redondance de la géolocalisation adresse IGN

L'appariement syntaxique est une phase longue, nécessitant dans l'absolue de comparer chaque adresse d'une base avec toutes les adresses de l'autre base pour établir la meilleure correspondance. Différentes optimisations permettent de limiter le temps de calcul. En particulier chaque commune est traitée séparément.

L'appariement syntaxique est caractérisé par une « distance syntaxique [1] ». L'appariement est recherché en plusieurs étapes de complexité croissante.

Dans l'ordre les méthodes successivement utilisées pour l'appariement syntaxique sont :

- exploitation d'un tampon permettant de tester immédiatement si la syntaxe de l'adresse est similaire à l'adresse précédente.
- sinon, recherche itérative limitée en cherchant l'adresse à apparier dans les enregistrements suivants l'adresse précédemment repérée.
- sinon, recherche directe de la correspondance parfaite au sein des adresses appartenant à la même section cadastrale. Cette dernière est tamponnée par une marge de sécurité de 100m de manière à ne pas exclure les adresses frontalières.
- sinon, recherche du meilleur correspondant au sein d'une liste d'adresse de même section cadastrale, de même préfixe de voie (rue, boulevard, place, etc…) et de même mot le plus long.
- sinon, recherche du meilleur correspondant au sein d'une liste d'adresse de même section cadastrale et de même préfixe de voie (rue, boulevard, place, etc…).
- sinon, recherche du meilleur correspondant au sein d'une liste d'adresse de même section cadastrale et de préfixe différents (la différence de préfixe étant comptée dans la distance syntaxique).
- sinon, recherche du meilleur correspondant au sein des adresses de toute la commune, de même préfixe de voie et de même mot le plus long.
- sinon, recherche du meilleur correspondant au sein des adresses de toute la commune et de même préfixe de voie.
- sinon, recherche du meilleur correspondant au sein des adresses de toute la commune et de préfixes différents.

E/ Définition d'un point de recherche médian

Il s'agit d'une phase de préparation à l'étape suivante.

La recherche des bâtiments d'accueil les plus probables présente trois justifications principales :

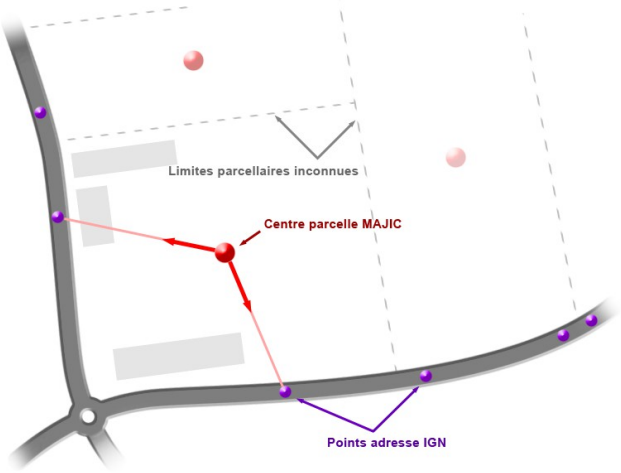
a) une parcelle pouvant présenter plusieurs adresses, l'objectif est de pouvoir « éclater » un géolocalisant unique (celui de la parcelle) en plusieurs géolocalisants, propres à chaque adresse (ceux des bâtiments probables d'accueil au sein de la parcelle).

b) sur des grandes parcelles, particulièrement en zone peu dense, le bâtiment d'accueil peut être éloigné du centre de la parcelle. Ainsi la précision de géolocalisation peut être accrue de plusieurs centaines de mètres dans ce cas.

c) fondamentalement, l'objectif est de tenter d'approcher la compatibilité la plus grande entre cette nouvelle couche informative et les autres couches SIG de référence (en particulier le RGE de l'IGN). À ce titre, tenter de repositionner la population et les entreprises au droit des bâtiments (couche BATI de l'IGN) est justifiable.

Le point de recherche médian a pour objectif d'optimiser la recherche des bâtiments d'accueil.

Tout d'abord, l'association des géolocalisants « adresse » et « parcelle » permet, dans le cas où la parcelle présenterait plusieurs adresses, de se rapprocher d'emblée de la zone de la parcelle la plus probable.



L'objectif est ensuite de favoriser la recherche des bâtiments les plus probables, sachant que le contour de la parcelle est dans notre cas inconnu. Seule la surface fiscale de la parcelle est connue.

On constate que les parcelles ne sont que très rarement allongées au-delà d'un rapport 10 (Longueur/Largeur). Cette constatation permet de définir un rayon de recherche maximal en fonction de la surface de la parcelle.

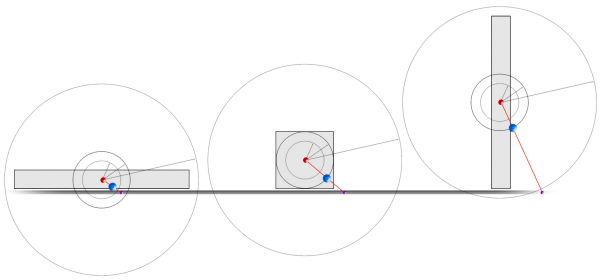
Par ailleurs, on considère que le localisant parcelle (quand il existe) est potentiellement plus fiable et plus proche du bâtiment d'accueil que le géolocalisant adresse. Ce dernier point est abusif en site dense (bâtiment en bord de rue sur des parcelles perpendiculaires à la voie et avec une grande densité de point adresse, mais, la densité parcellaire est alors telle que la précision recherchée, de l'ordre de quelques dizaines de mètres, est garantie dans tous les cas).

Deux rayons spécifiques sont calculés :

Rayon Proche = $\sqrt{S_{Parcelle}}/2$ *(Rayon garantissant que le point médian à de forte chance de rester proche de la parcelle)*

Rayon Maximum = Min (1500m , 1.6 x (Max($\sqrt{S_{Parcelle}}$,100))) *(Rayon limite à partir duquel tout bâtiment sera écarté de la recherche. Afin de corriger d'éventuelle erreur de la BDD majic la surface des parcelles est majorée par une valeur de 100m². Par ailleurs le rayon de recherche est borné par une valeur de 1500m).*

Le point médian est choisi au milieu du segment formé par les géolocalisants parcelle MAJIC et adresse IGN tant que sa distance au centre de la parcelle est inférieure au « rayon proche », et placé à la limite du « rayon proche » dans les autres cas.



En cas d'absence de point adresse IGN, le point médian est équivalent au géolocalisant parcelle.

En cas d'absence de point parcelle MAJIC, le point médian est équivalent au géolocalisant IGN à condition qu'il soit de redondance faible. La redondance est égale au nombre de points adresse co-géolocalisés au point considéré dans la BDD Adresse de l'IGN. Au-delà d'un certain niveau de redondance (30 pour l'instant), la géolocalisation est jugée trop peu fiable.

F/ Recherche des bâtiments d'accueil probables

Pour chaque point médian sont ensuite recherchés les bâtiments d'accueil les plus probables au sein de la couche BATI de la BD TOPO de l'IGN.

Sont pris en compte les bâtiments « indifférenciés », « remarquables » et « industriels ».

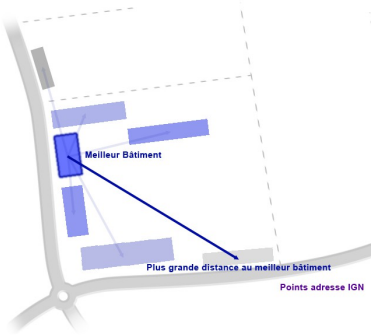
L'ensemble des bâtiments de la couche BATI de l'IGN situés dans un cercle centré au « point médian » et de rayon égal au « rayon maximal » est extrait.

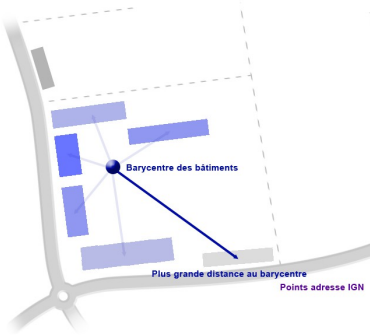
Les locaux comptabilisés à cette adresse sont ensuite distribués par partie entière sur les bâtiments identifiés, en fonction de leur surface d'accueil et de leur distance au point médian. La loi de distribution est en SurfaceBatiment / DistancePoint médian ^{2} .

Un diagnostic est ensuite fourni à partir du sous-ensemble de bâtiment ayant reçu des locaux.

Le meilleur bâtiment (celui ayant reçu le plus de locaux) est identifié, ces coordonnées géographiques sont extraites et la distance maximale de ce bâtiment aux autres bâtiments ayant reçu des locaux à cette adresse est calculée. Le barycentre (où chaque bâtiment est pondéré par le nombre de local reçu) du groupe de bâtiment d'accueil est calculé ainsi que la distance de ce point à l'ensemble des bâtiments ayant reçu des locaux.

Ainsi, l'erreur potentielle liée à l'identification du bâtiment d'accueil est systématiquement qualifiée.

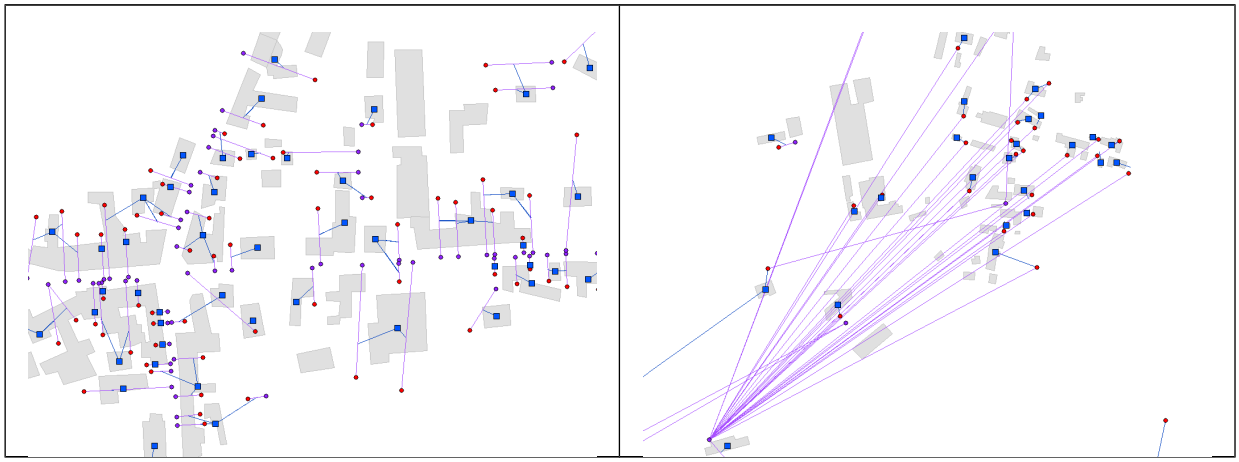




Identification des bâtiments autour du point médian	Identification du bâtiment recevant le plus de locaux et calcul de la distance maximale aux autres bâtiments identifiés	Calcul du barycentre et calcul de la distance maximale aux autres bâtiments identifiés

Si aucun bâtiment situé à une distance inférieure au « rayon maximal » n'est trouvé, on conserve néanmoins les coordonnées du point médian (supposé statistiquement plus performant que le point parcelle initial). Les distances au barycentre et au meilleur bâtiment sont remplacées dans ce cas par le rayon de recherche maximal.

Extrait des résultats sur le département 44 :



Exemple en zone dense :
Densité des BD MAJIC et IGN équivalentes. Le résultat est de même densité mais avec recentrage sur les bâtiments

Exemple en zone peu dense :
En cas de regroupement des localisations dans la BD Adresse IGN, la méthode compense par la recherche des bâtiments occupés

● Point Adresse IGN ● Point Parcelle MAJIC ■ Meilleur Bâtiment identifié

La structure de la table résultante est la suivante :

BDD des locaux fiscalisés géolocalisés :

► Nombre de locaux hab.
► Nombre de locaux pro.
► Surface fiscale de la parcelle d'appartenance ($S_{Parcelle}$)
► Surface fiscale bâtie agrégée ($S_{Bat\ MAJIC}$)
► Adresse postale
► Géolocalisant de la parcelle d'appartenance ($X,Y_{Parcelle\ MAJIC}$)
► Géolocalisant du point adresse IGN apparié ($X,Y_{Adresse\ IGN}$)
► Distance syntaxique IGN-MAJIC
► Niveau de redondance de la géolocalisation adresse IGN
► Géolocalisant du point médian ($X,Y_{Médian}$)
► Nombre de bâtiment recevant des locaux à cette adresse
► Géolocalisant du meilleur bâtiment d'accueil ($X,Y_{Best\ Batiment\ IGN}$)
► Distance maximale au meilleur bâtiment d'accueil
► Géolocalisant du barycentre des bâtiments ($X,Y_{Bary\ Batiment\ IGN}$)
► Distance maximale au barycentre

1 La distance syntaxique rapide s'apparente à une méthode de « Levenshtein », un peu moins précise mais plus rapide à calculer. Elle est explicitée en annexe 2.

2 La surface d'accueil est égale au nombre d'étage supposé multiplié par la surface au sol du bâtiment, telle que figurant dans la couche BATI de la BD Topo. La méthode de calcul est explicitée en annexe 3.

ANNEXE 2 : Méthode de calcul de la distance syntaxique

Afin de réaliser des appariements d'adresses en masse il est nécessaire de pouvoir calculer l'écart entre deux syntaxes d'adresse et de manière très rapide.

Les méthodes complexes dites de « distance phonétique » ne peuvent être ici utilisées en raison de leur lenteur.

Une méthode classique d'évaluation de la distance syntaxique est celle de « Levenshtein ». Elle évalue une distance syntaxique égale au nombre minimal de caractères qu'il faut supprimer, insérer ou remplacer pour passer d'une chaîne de caractères à l'autre. Il s'agit donc d'une méthode de comparaison ordonnée scrutant les défauts syntaxiques à tous les niveaux de la chaîne de caractères décrivant l'adresse.

La comparaison brutale d'une adresse complète à une autre par la méthode de Levenshtein a été estimée trop lente pour s'appliquer à des BDD France entière.

Statistiquement, il a été vérifié que l'essentiel des écarts syntaxiques entre la base MAJIC et celle de l'IGN correspondent aux cas suivants :

- différence d'abréviation
- ordre des mots
- faute d'orthographe (ou réduction) essentiellement en fin de mot.

Tenant compte de ces critères, les étapes de la technique de calcul de la distance syntaxique rapide utilisée sont les suivantes :

PRE-TRAITEMENT

- harmonisation des abréviations en exploitant une table de correspondance
- extraction du préfixe de voie si reconnu
- séparation des mots restants.

TRAITEMENT

- pour chaque mot de la première adresse, recherche du mot de la deuxième adresse le plus proche (minimisation de la distance syntaxique rapide pour chaque mot)
- la distance syntaxique globale est la somme des distances syntaxiques rapides trouvées pour chaque mot.

La distance syntaxique rapide mot à mot, supposant que les erreurs les plus probables sont en fin de mot, compare les lettres de deux mots de manière séquentielle (sans saut, contrairement à Levenshtein).

Pour donner un ordre d'idée, là où la comparaison de deux chaînes de caractères via Levenshtein est fonction de la longueur de la plus longue au carré, la méthode rapide séquentielle est uniquement liée à la longueur.

La comparaison des chaînes de caractères par mot entier permet ensuite de limiter le temps de calcul car elle est fonction du nombre de mots et non du nombre de caractères global.

ANNEXE 3 : Méthode de distribution des locaux

Les parcelles vectorisées n'étant pas exploitées dans la présente méthode, l'identification des bâtiments probables pouvant correspondre à une adresse donnée est forcément imprécise.

Il n'est pas possible de savoir exactement, particulièrement si une parcelle est très allongée ou biscornue, si un bâtiment appartient bien à la parcelle en cours de traitement.

Par ailleurs, la notion de bâtiment construit n'existant pas directement dans la base MAJIC, même au sein d'une parcelle bien identifiée, il ne serait pas possible d'attribuer rigoureusement les locaux comptabilisés à tel ou tel bâtiment réel.

Une règle de distribution entière des locaux au sein des bâtiments probables, pour chaque adresse, est donc utilisée.

À partir du point médian (servant de centre à l'identification des bâtiments probables), les locaux sont distribués en fonction de la capacité et de la probabilité de chaque bâtiment :

- la capacité est associée à la surface totale estimée du bâtiment
- la probabilité est liée à l'inverse de la distance au point médian.

La règle de distribution retenue est fonction de : Surface totale du bâtiment / Distance au point médian au carré.

La surface totale du bâtiment est estimée à partir des seules informations de la couche BATI de la BD-TOPO de l'IGN.

La surface au sol est assimilée à la surface du polygone figurant chaque bâtiment.

Le nombre d'étage est estimé à partir d'une loi statistique exploitant la seule hauteur du bâtiment :

- Si hauteur<6 Nb étage = 1
- Si 6≤hauteur<9.3 Nb étage = partie entière (hauteur/3)
- Si 9.3≤hauteur Nb étage = partie entière ((hauteur-3.5)/2.9) + 1

La surface totale estimée est assimilée à : Surface au sol x Nombre d'étages estimé.

ANNEXE 4 : Optimisation Software / Hardware

La production de la BDD est réalisée sur le logiciel MapInfo, par l'intermédiaire de scripts MapBasic.

En raison du périmètre national du projet, ainsi que de la relative complexité des données et traitements, la production de la BDD Géolocaux est relativement longue (plusieurs semaines à plusieurs mois selon les moyens informatiques mis en œuvre).

Pour cette raison des développements spécifiques ont été réalisés pour utiliser MapInfo en mode parallélisé de manière à exploiter pleinement les capacités des machines actuelles.

L'utilisation parallélisée de Mapinfo induit par ailleurs une multiplication des accès disques synchrones (en entrée et en sortie). Ceci peut provoquer de fortes chutes de performance par encombrement du bus chargé des échanges entre la mémoire vive et le disque dur de l'ordinateur.

Pour remédier à ce goulot d'étranglement hardware, la plupart des opérations de lecture et écriture de MapInfo (correspondant aux sauvegardes des étapes transitoires) sont déportées sur un disque virtuel (en RAM), alors que les données sources et résultats restent stockées sur le disque dur.

Une console pilote les satellites dédiés au traitement des données et gère la liste des taches à effectuer en parallèle. Chaque tâche est relative à une étape de traitement pour un département.

À chaque fois, que la console détecte qu'un cœur de calcul se libère, elle lance un nouveau satellite pour effectuer la tache suivante de sa liste de manière à exploiter l'intégralité de la puissance de calcul disponible.

Chaque satellite de traitement est associé à une instance de MapInfo lancé en mode serveur. Ce mode permet de lancer MapInfo sans interface afin d'économiser de la mémoire vive et d'éviter toute manipulation intempestive, s'agissant de traitements totalement automatisés.

Pour information, le traitement de l'intégralité de la BDD Géolocaux sur une machine bi-hexa-cores prend environ 2 semaines.

ANNEXE 5 : Contrôles qualité de la BDD Géolocaux

Contrôle externe

La différence moyenne en % du nombre de foyer retenus entre le nombre de foyer de la base locaux (BDD_GEOLOCAUX) et le nombre de logements de la base INSEE (2009) est de 3,44 %.

Contrôle interne (entre base locaux et base MAJIC)

- Le pourcentage moyen de perte de foyer est de 0,24 %
- Le pourcentage moyen de perte de locaux professionnels est de 0,35 %
- La moyenne des rayons moyens de recherche des bâtiments est de 46,57 m
- La moyenne des écarts type des rayons de recherche est de 55,37 m
- Pour un rayon de recherche de 50,00 m moyen, il y a 66 % des foyers géolocalisés

Rayon de recherche du bâtiment (m)	50 m	100 m	150 m	200 m	250 m
% de foyers géolocalisés	66 %	86 %	93 %	96 %	98 %

Un tableau d'analyse a été réalisé pour fournir des informations statistiques complémentaires, au niveau départemental.

Les DOM ne sont pas intégrés au contrôle, le traitement étant spécifique.

--

Ainsi la base est pertinente pour toute utilisation dans le cadre de ses objectifs premiers :

- cartographier à l'échelle d'un quartier la couverture par les services de communications électroniques (échelle de la planification des déploiements),
- calculer des taux de couverture de ces services (à l'échelle communale, du quartier, des hameaux, etc.) : la méthode conserve le dénombrement des locaux au sein de ses polygones regroupant un même niveau de service.

En revanche, cette base ne peut en aucun cas être utilisée pour déterminer une éligibilité à la parcelle, ou au bâtiment. Toute utilisation autre offerte par cette nouvelle base devra faire l'objet de sa propre analyse de pertinence.

Centre d'études et d'expertise sur les risques, l'environnement, la mobilité et l'aménagement
Point d'Appui National – Aménagement Numérique des Territoires (PAN ANT)
www.ant.developpement-durable.gouv.fr - ant.dvt.DterOuest@cerema.fr
Direction territoriale Ouest : MAN – 9 rue Viviani – BP 46223 – 44262 Nantes cedex – Tél : +33(0)2 40 12 83 01
Siège social : Cité des Mobilités - 25, avenue François Mitterrand - CS 92 803 - F-69674 Bron Cedex - Tél : +33 (0)4 72 14 30 30
Établissement public : Siret 130 018 310 00 222 www.cerema.fr