Quelques exemples d'utilisation de données libres en sciences sociales et économiques

Journée d'étude PROGEDO

Thibault Laurent

Toulouse School of Economics, CNRS

27 Juin 2017

Des exemples de projets à TSE...

Etude sur la prédiction de l'usage des sols

Travaux d'ingénierie électorale

... utilisant des données non libres

1. Groupe “économie de l’alimentation” : impact des politiques publiques sur la consommation alimentaire
   - Données *Kantar WorldPanel* : actes d’achat alimentaire d’un panel de ménages.
   - Méthodes : outils variés en économétrie de la demande (ex : modèles de choix discrets à utilité aléatoire).
   - Logiciels utilisés : Stata, Matlab (non libres); R, Python (libres).
   - Données croisées avec d’autres souces de données, parfois libres (par exemple Ciqual, ANSES).
2. Groupe “statistique et mathématique de la décision” : analyse des flux de transports aériens (avec Airbus)
   - Données *Sabre Airline Solutions proprietary data set*, sur le trafic aérien à l’échelle mondiale.
   - Modèle d’économétrie spatiale.
   - Données croisées avec d’autres sources.

... mélangeant des données libres et propriétaires

Labex IAST (économie, biologie, droit, sciences politiques, etc.)

1. Effet de la médiatisation sur les décisions de justice (Arnaud Philippe).
   - Ministère de la justice : données (sur les détenus) consultables uniquement dans les salles du ministère.
   - Données libres sur les crimes et délits enregistrés par la police nationale (par département) : `https://www.data.gouv.fr/`
   - Enquête cadre de vie et sécurité : données produites par l'INSEE, disponibles via le réseau Quetelet.
2. Expliquer les différences hommes/femmes sur le marché du travail (Marie Lalanne).
   - Base de données *BoardEx Lt* : croisement de données sur les CV et entreprises (échelle européenne).
   - EDGAR Data : entreprises américaines côtées en bourses.

... produisant des données

Groupe "économie expérimentale" et IAST : analyse du comportement d'individus soumis à certaines expériences

1. Intelligence collective (Adrien Blanchet). Exemple : une matrice de taile 12×12 avec des nombres cachées dans chaque cellule. Objectif : trouver le nombre le plus grand avec un nombre limité d'essais, seul ou à plusieurs.
2. Jeu de confiance (Marie Lalanne). Objectif : hommes et femmes forment-ils leurs réseaux sociaux de la même façon ?

Les jeux de données peuvent ensuite être rendus acessibles via les revues scientifiques (par exemple, CSBIGS, `http://publications-sfds.fr/index.php/csbigs`).

Enseignement et données libres

1. Master 1 filière économétrie, statistique (Christine Thomas-Agnan)
   - Open data Toulouse : `https://data.toulouse-metropole.fr/pages/accueil/`
   - `https://opendata.socrata.com/` : comptes des partis politiques français 2004-2009.
   - `https://www.data.gouv.fr/fr/datasets` : insertion professionnelle des diplômés de Master en universités et établissements assimilés, balances comptables des communes.
2. M2 statistique et économétrie (Anne Ruiz-Gazen) : concours "Data science" avec d'autres universités
   - Challenge "Open Bike" : données sur les vélo Toulouse. Objectif : prédire le nombre de vélos disponibles sur certaines stations à un temps t.
   - Systèmes de recommandation par filtrage collaboratif pour le commerce en ligne. Kaggle data set : données sur la base de données "movieLens".

Des exemples de projets à TSE...

Etude sur la prédiction de l'usage des sols

Travaux d'ingénierie électorale

ANR ModuLand (2011–2016)

- Thématique : modélisation économétrique et statistique des usages des sols et l'étude de leurs impacts sur l'environnement (gaz à effet de serre, biodiversité, qualité de l'eau).
- Exemples :
  - Impact de la rente foncière sur l'usage des sols,
  - Impact économique du changement climatique sur l'agriculture,
  - Déterminants de l'adoption de mesures agro-environnementales (conversion au bio, diversification des cultures).
- Article présenté ici :

Chakir R., Laurent T, Ruiz-Gazen A., Thomas-Agnan C. et Vignes C. (2017). Prédiction de l'usage des sols sur un zonage régulier à différentes résolutions et à partir de covariables facilement accessibles. *Revue économique*, **68**.

Objectifs

Deux objectifs :

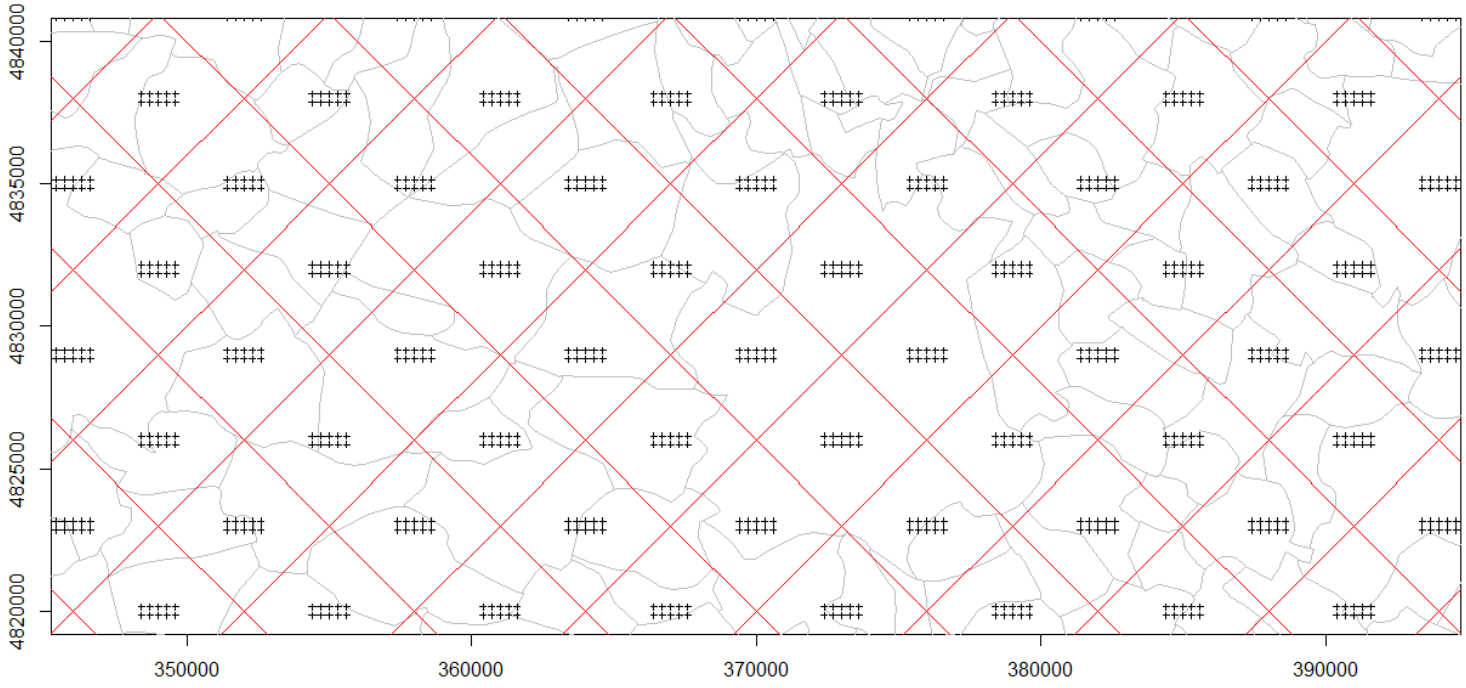
1. Prédire l'usage des sols dans la région Midi-Pyrénées (en 5 catégories) en utilisant des données facilement accessibles.
2. Comparer les mesures de prédictions selon l'échelle spatiale (du point jusqu'à des carreaux réguliers de plus en plus grand).

On s'intéresse à deux critères de comparaison :

- Le taux de bien classés à l'échelle du point.
- Le score de Brier à l'échelle du carreau (erreur quadratique moyenne des proportions estimées).

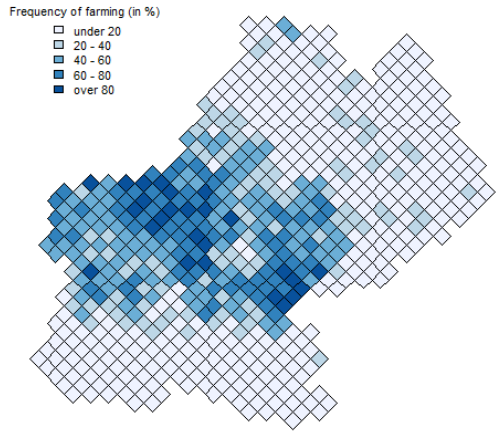
La variable à expliquer (données non libre)

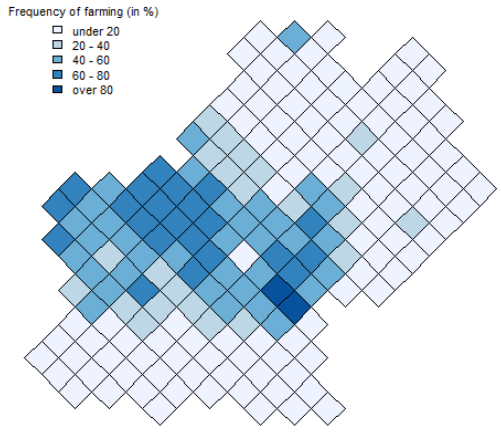
- Données Teruti-Lucas : 25317 points en Midi-Pyrénées.
- 5 usages : urbain, agricole, forêts, prairies, sans usage.

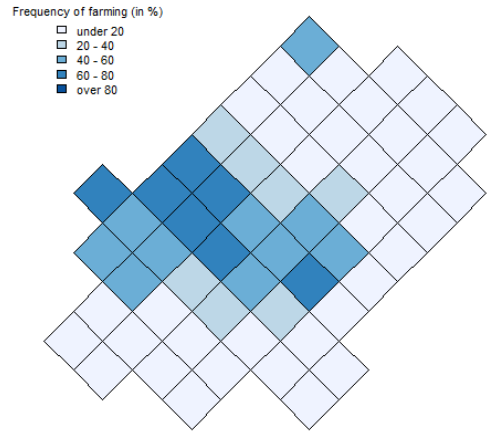


- Source : Service de la Statistique et de la Prospective, Ministère de l'Agriculture.

La variable à expliquer à différents niveaux de résolutions

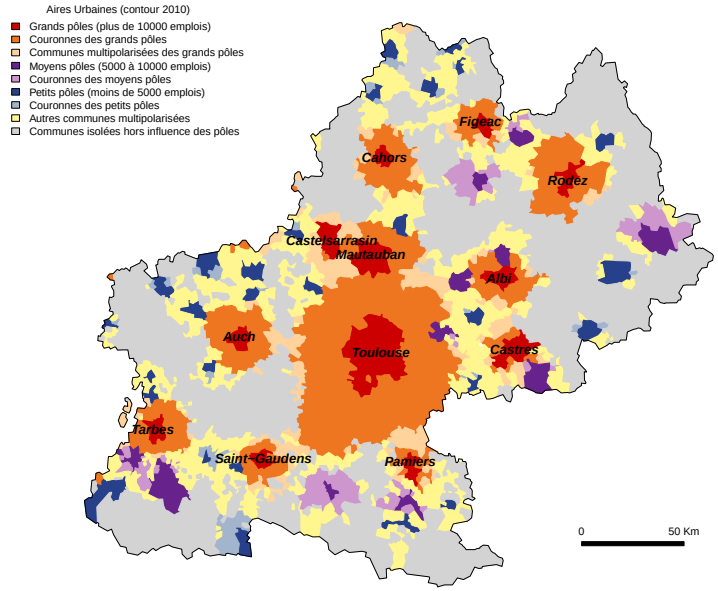






Les fonds de carte (données libres)

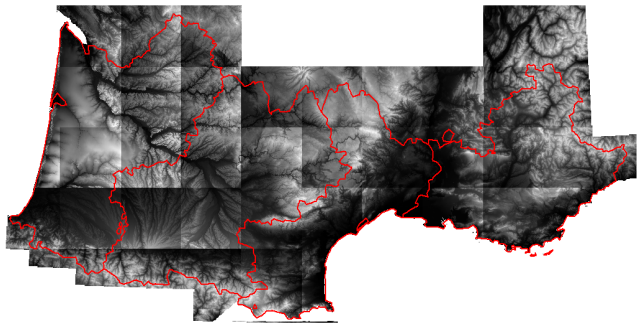
- IGN : contours des limites administratives françaises (régions, départements, arrondissements, communes) : `http://professionnels.ign.fr/geofla`
- Contours de limites administratives à l'échelle mondiale : `http://www.gadm.org/`



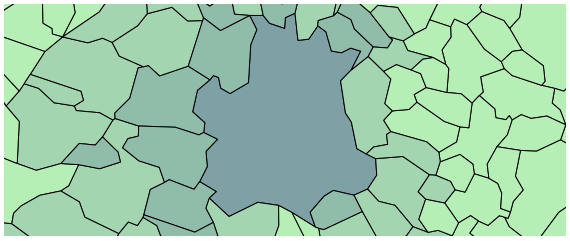
Source : INSEE (`https://www.insee.fr/fr/information/2115011`)

Autres variables fournies par l'IGN et l'INSEE (données libres)

- Base de données "BD ALTI" (IGN) : altitude à des pas de $250m$.

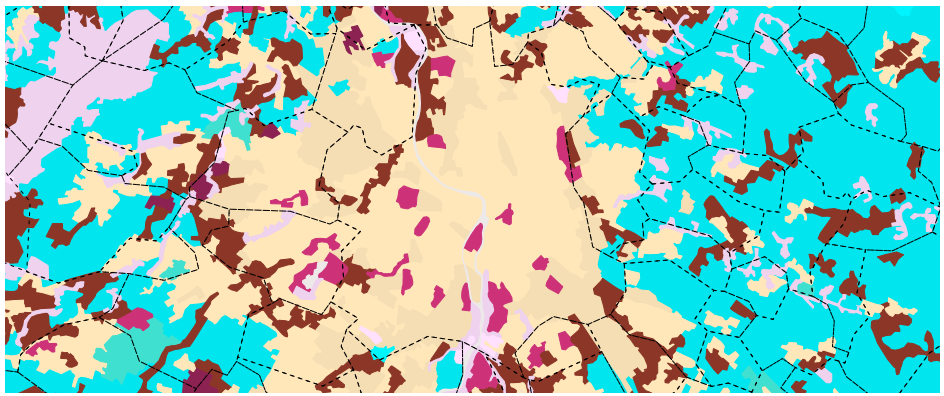


- Au niveau communal (INSEE), densité de population, part d'agriculteurs, part de cadres, grands pôles urbains.



Corine Land Cover 2006 (données libres)

- Agence européenne pour l'environnement
- Service de l'observation et des statistiques du ministère chargé de l'environnement : `http://www.statistiques.developpement-durable.gouv.fr/clc/fichiers/`

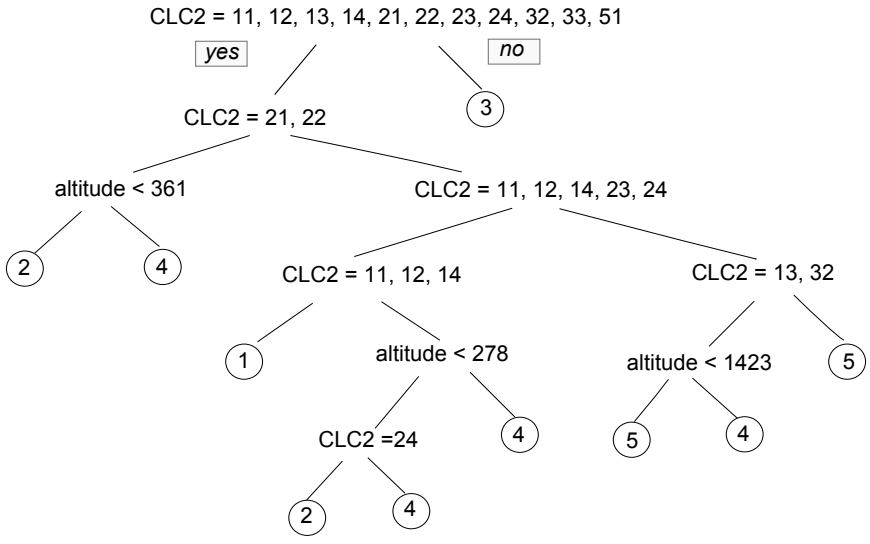


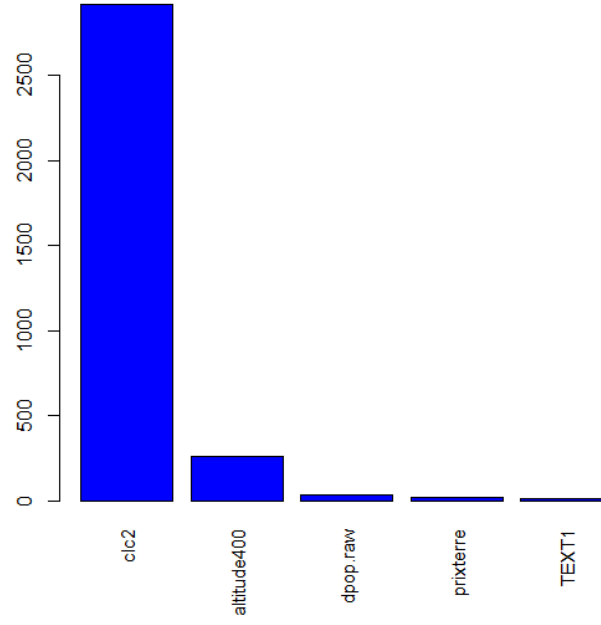
Occupation des sols (niveau 2) : zones urbanisées, zones industrielles ou commerciales, terres arables, cultures permanentes, forêts et milieux semi-naturels, etc.

Autres données utilisées

- Données “Agri4cast interpolated meteorological data” (Institute for Environment and Sustainability), libres sur demande. Température minimum, température maximum, température moyenne, somme annuelle des précipations, vitesse moyenne du vent. Echelle : grille 25×25 km.
- Données GISSOL : données géographique des sols de France (INRA Orléans). Echelle géographique : NRA (nouvelles régions agricoles). Données non libres

Arbre de classification et importance des variables

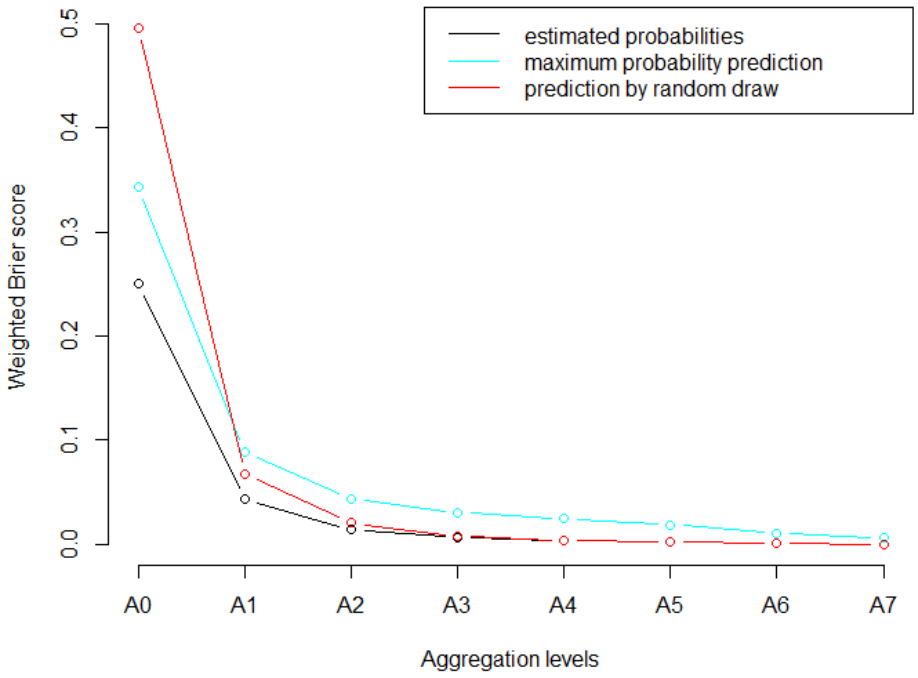




Pourcentage de bien classés : 65.25%

Estimation au niveau des carreaux

A partir des résultats au niveau du point, estimer l'usage des sols au niveau du carreau



Aggrégation au niveau des segments (A_1) semble suffisante pour améliorer les résultats.

Des exemples de projets à TSE...

Etude sur la prédiction de l'usage des sols

Travaux d'ingénierie électorale

Données sur les élections

1. Données réelles :
   - Résultats d'élections sur la France entière `https://www.data.gouv.fr` (au niveau régions, départements, cantons, circonscriptions législatives ou communes), à l'échelle des bureaux de vote de certaines villes comme Toulouse `https://data.toulouse-metropole.fr/`.
   - Résultats des élections américaines `https://www.data.gov/`
2. Données expérimentales pour mesurer l'importance du mode de scrutin sur une élection (Karine Van der Straeten).

Projets en cours

1. Olivier de Mouzon, Thibault Laurent, Dominique Lepelley and Michel Le Breton (2017). The Theoretical Shapley-Shubik Probability of an Election Inversion in a Toy Symmetric Version of the U.S. Presidential Electoral System. *TSE Working Paper*.
2. Olivier de Mouzon, Thibault Laurent, Dominique Lepelley and Michel Le Breton (2017). Exploring the Effects on the Electoral College of a Regional Popular Vote Interstate Compact.

Importance du mode de scrutin

Retour sur le 1er tour de l'élection présidentielle 2017. On change le mode scrutin :

1. On choisit un découpage administratif (département ou circonscription législative)
2. Le parti avec la plus grande majorité dans une zone gagne 1 grand électeur
3. Le parti avec le plus grand nombre de grands électeurs remporte l'élection

- Découpage électoral = département. Nombre de grands électeurs : FILLON 9, LE PEN 47 (W), MACRON 43, MÉLENCHON 7
- Découpage électoral = circonscriptions législatives. Nombre de grands électeurs : FILLON 53, LE PEN 216, MACRON 230 (W), MÉLENCHON 67