AVIGNON UNIVERSITÉ

M2 ILSEN – 2023/24
UE Business intelligence – Systèmes décisionnels
ECUE Application Business Intelligence

Vincent Labatut

Projet | Données ouvertes des marchés publics

1 Présentation

Contexte. L'expression *marchés publics* désigne le fait que des institutions publiques achètent des travaux, fournitures, et services, auprès d'opérateurs privés (ou publics). Il s'agit d'un secteur économique essentiel, puisqu'au sein de l'UE ces transactions représentent 15% du PIB. La procédure d'attribution d'un marché public est encadrée par la loi, et suit une procédure bien précise. En particulier, lorsque le montant estimé du marché dépasse un certain seuil, cette procédure oblige l'acheteur à publier l'appel d'offres ainsi que l'avis d'attribution correspondant, via des médias spécialisés. Pour cette raison, l'activité relative aux marchés publics génère une grande quantité de données.

Données. Le projet DeCoMaP [1] (*Détection de la Corruption dans les Marchés Publics*) est financé par l'ANR [2] (*Agence Nationale de la Recherche*) et porté par Avignon Université. Il vise, entre autres, à proposer des méthodes automatiques de détection de fraude dans les marchés publics. La base de données FOPPA [3] (*French Open Public Procurement Award notices*) a été conçue dans le cadre de ce projet. Elle contient les appels d'offres et les avis d'attribution des lots concernant les marchés publics en France pour la période 2010–2020.

Objectif. L'objectif de ce projet est de mener une analyse descriptive de ces données, afin de répondre à différentes questions détaillées plus loin. La nature de cette analyse est assez large, le projet vous laisse une grande liberté sur ce point.

2 Données

Origine. La FOPPA est basée sur des données brutes issues de la TED [4] (*Tenders Electronic Daily*), une base de données européenne. Ces données sont caractérisées par la présence d'un grand nombre problèmes (données erronées, manquantes, mal codées...), dont les principaux ont été traités pour produire la FOPPA. Il subsiste néanmoins un nombre significatif de problèmes dans les données ainsi obtenues. Le traitement utilisé pour produire la FOPPA est décrit en détail dans un rapport technique [2], et résumé dans un article [3]. Le rapport technique inclut des explications sur le fonctionnement des marchés publics, destinées à des non-spécialistes, et qui peuvent vous être utiles pour comprendre le fonctionnement de ce système, et donc faciliter l'analyse des données et l'interprétation des résultats afférents.

Concrètement, la FOPPA est disponible sous la forme de tables de données CSV ainsi que de dump SQL.

Structure. Les Tables 1–5 listent leurs attributs ainsi que la signification de ceux-ci. L'architecture de la base de données est décrite dans la Figure 1.

Il apparaît clairement dans cette figure que la table de données principale est `Lots`. Celle-ci représente les lots mis sur le marché lors de la période couverte par la FOPPA. Les nombreux champs qui la constituent décrivent différents aspects du lot, mais aussi de la procédure d'attribution, et du résultat de l'attribution. Ils permettent également de relier le

1. https://anr.fr/Projet-ANR-19-CE38-0004
2. https://anr.fr/
3. https://doi.org/10.5281/zenodo.7808664
4. https://ted.europa.eu/

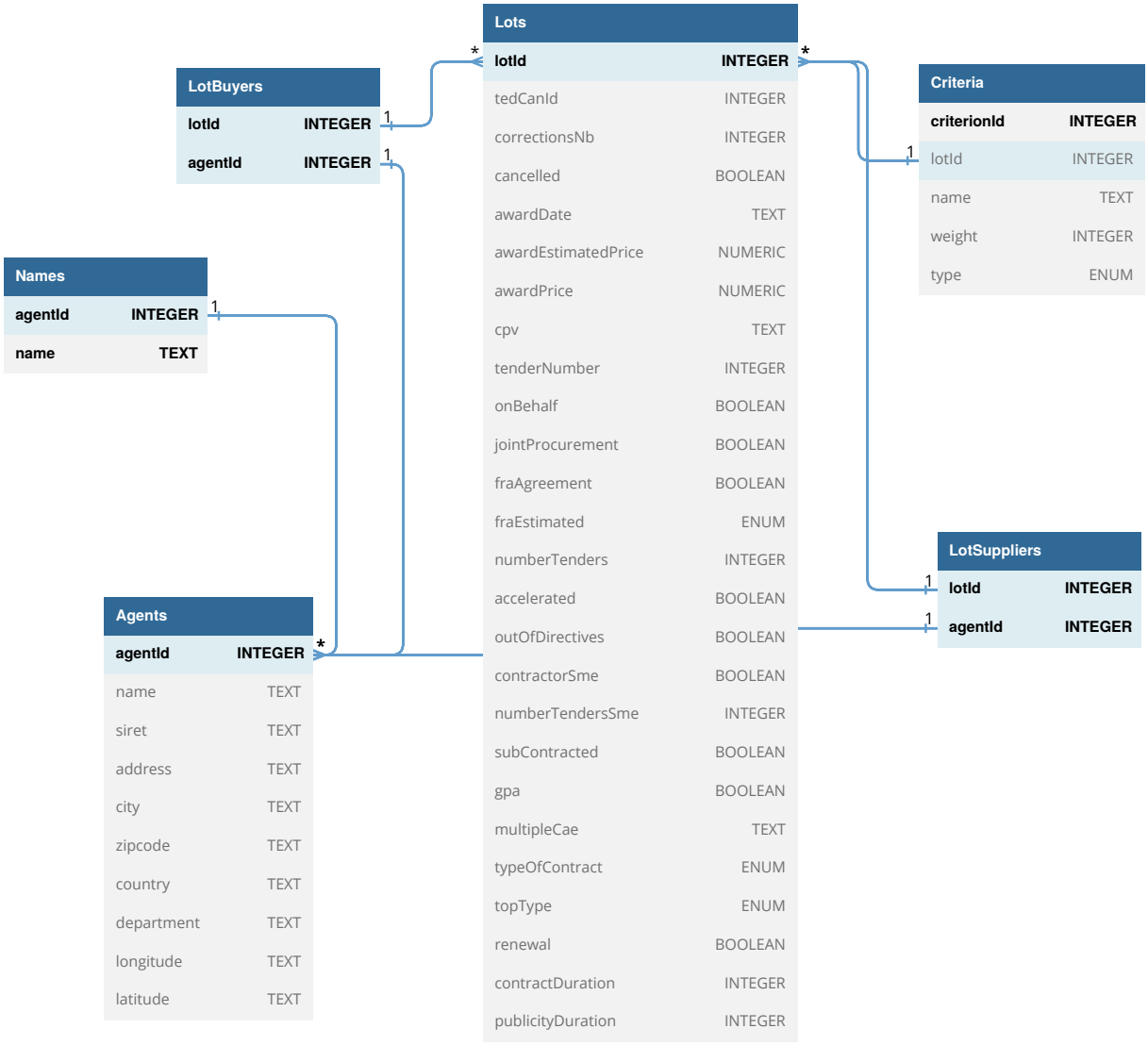

Figure 1. Architecture de la base de données FOPPA.

lot aux acheteurs et fournisseurs concernés. Ceux-ci sont représentés dans la table `Agents`, de façon indifférenciée car un agent économique peut très bien jouer le rôle d'*acheteur* pour un lot donné et de *fournisseur* pour un autre lot. Les tables `LotBuyers` et `LotSuppliers` sont nécessaires pour relier les acheteurs et fournisseurs aux lots les concernant, car il s'agit de relations multiples.

Un lot est attribué en prenant en compte plusieurs critères prédéfinis, ce qui requiert là aussi une association multiple. C'est la raison pour laquelle la table Criteria est dédiée au stockage des critères et de leurs caractéristiques. Enfin, un agent pouvant posséder plusieurs noms, la table `Agents` ne stocke que le principal, tandis que les noms secondaires sont listés dans la table `Names`.

Attribut	Signification
`agentId`	Clé primaire
`name`	Nom principal de l'agent (cf. Table 5)
`siret`	Numéro d'identification unique dans la base SIRENE
`address`	Voie et numéro dans l'adresse postale
`city`	Ville dans l'adresse postale
`zipcode`	Code postal dans l'adresse
`country`	État membre
`department`	Département français
`longitude`	Position spatiale (X)
`latitude`	Position spatiale (Y)

Table 1. Attributs de la table `Agents`, représentant les agents économiques jouant le rôle d'acheteur ou de fournisseur dans un marché public (voire plusieurs).

Attribut	Signification
`lotId`	Clé primaire
`tedCanId`	ID TED de l'avis d'attribution du lot
`correctionsNb`	Nombre de correctifs publiés pour ce lot
`cancelled`	Booléen indiquant si l'appel d'offres du lot a été annulé
`awardDate`	Date à laquelle la décision d'attribution a été prise pour ce lot
`awardEstimatedPrice`	Estimation du prix du lot effectuée dans l'appel d'offres
`awardPrice`	Prix effectif du lot indiqué dans l'avis d'attribution
`cpv`	Code *Common Procurement Vocabulary* du lot
`numberTenders`	Nombre d'offres effectuées pour le lot
`onBehalf`	Booléen indiquant que l'acheteur est un groupement
`jointProcurement`	Booléen indiquant s'il s'agit d'un marché conjoint
`fraAgreement`	Booléen indiquant si le lot fait partie d'un accord cadre
`fraEstimated`	Champ suggérant l'existence d'un accord cadre (le cas échéant)
`lotsNumber`	Nombre total de lots dans la notice d'attribution
`accelerated`	Booléen indiquant l'utilisation de la procédure rapide
`outOfDirectives`	Booléen indiquant un avis d'attribution sans appel d'offres associé
`contractorSme`	Booléen indiquant si le gagnant est une PME
`numberTendersSme`	Nombre d'offres issues de PME pour ce lot
`subContracted`	Booléen indiquant si le lot est sous-traité
`gpa`	Booléen indiquant un lien avec l'*Accord sur les Marchés Publics*
`multipleCae`	Booléen indiquant si l'avis d'attribution liste plusieurs acheteurs
`typeOfContract`	Contrat de fournitures (`S`), travaux (`W`), ou services (`U`)
`topType`	Type de procédure d'attribution
`renewal`	Possibilité de renouveler le contrat
`contractDuration`	Durée du contrat
`publicityDuration`	Durée de la période de publicité de l'appel d'offres

Table 2. Attributs de la table `Lots`, représentant les lots faisant l'objet des marchés publics.

Attribut	Signification
`lotId`	Clé étrangère indiquant le lot concerné
`agentId`	Clé étrangère désignant l'agent concerné

Table 3. Attributs des tables `LotsBuyers` et `LotsSuppliers`, qui permettent de relier les tables `Agents` et `Lots`, et ainsi d'implémenter les relations multiples correspondantes.

Attribut	Signification
`criterionId`	Clé primaire
`lotId`	Clé étrangère désignant le lot concerné
`name`	Représentation textuelle du critère
`weight`	Importance du critère dans le processus d'attribution
`type`	Catégorie du critère

Table 4. Attributs de la table `Criteria`, qui représente les critères d'attribution associés aux lots de la table `Lots`.

Attribut	Signification
`agentId`	Clé étrangère désignant l'agent concerné
`name`	Nom (secondaire) associé à cet agent

Table 5. Attributs de la table `Names`, qui contient les noms secondaires attribués aux agents économiques de la table `Agents`, en plus de leur nom principal (cf. Table 1).

3 Préparation

Exploration. Dans un premier temps, naviguez manuellement dans les données afin de mieux les appréhender. Vous devez d'abord étudier chaque variable prise individuellement :

- Identifiez la *nature* de la variable, ainsi que son *codage*. Discutez-les, notamment si vous jugez que le codage n'est pas approprié.
- Produisez un graphique montrant la distribution de la variable, et discutez-le. S'agit-il d'une distribution standard, et si oui laquelle ?
- Calculez les statistiques standard, en fonction de la nature de la variable : moyenne, écart-type, quantiles, mode, min, max, etc. Discutez-les.
- Cherchez d'éventuels problèmes : valeurs manquantes ? Valeurs aberrantes ? Valeurs erronées ?

Une fois chaque variable traitée individuellement, utilisez ces différentes informations pour comparer celles pour lesquelles une telle comparaison est pertinente.

Puis, étudiez les possibles associations statistiques pour toutes les paires de variables au sein de la même table de données :

- Produisez un graphique montrant les deux variables, en choisissant un type de graphique adapté aux natures de ces deux variables. Discutez-le.
- Calculez une mesure d'association entre ces deux variables. Discutez-la, notamment par rapport au graphique.

Intéressez-vous aux paires de variables fortement associées, ainsi qu'aux associations ou absences d'association que vous jugez inattendues. Utilisez ces résultats pour approfondir votre compréhension des données, et éventuellement pour détecter d'autres problèmes.

Nettoyage. L'étape suivante est celle du nettoyage des données, qui consiste à traiter les différents problèmes que vous avez identifiés. Certains résultats produits au cours de l'exploration permettent de prendre des décisions pour ce qui concerne une partie de ces problèmes. En effet, l'exploration a déjà permis d'identifier des erreurs au niveau des variables individuelles (codage inapproprié, valeurs manquantes, aberrantes, erronées), et l'étude d'association a permis d'identifier des paires de variables qui seraient redondantes.

La question est alors de savoir comment résoudre ces différents problèmes. Il y a essentiellement deux approches : substitution ou suppression. On peut supprimer des entrées ou des attributs, mais bien sûr cela aboutit à une perte d'information. On peut substituer aux valeurs manquantes, erronées ou aberrantes, des valeurs neutres, mais avec un risque de biaiser les données. On peut aussi effectuer la substitution en exploitant une source de données

externe [5] qui vient compléter les nôtres. Vous devez considérer les différentes méthodes à votre disposition, et évaluer lesquelles sont les plus appropriées pour traiter les différents cas que vous avez identifiés. Surtout, il est important de bien justifier ces choix méthodologiques dans votre rapport. Et tout cela nécessite de bien comprendre les données.

4 Analyse

Questionnements. Une fois les données explorées et nettoyées, il est possible de se concentrer sur certaines questions de recherche. L'objectif est alors de proposer des méthodes qui permettront de répondre à ces questions. Voici quelques exemples de questions, dont vous pouvez vous saisir, mais vous pouvez également en proposer d'autres.

Peut-on identifier des *classes de similarité* d'agents économiques ? De lots ? Quelles sont les variables et valeurs qui permettent de distinguer ces classes, et/ou qui sont caractéristiques de certaines classes ? Comment interpréter ces classes, que signifient-elles ? Y a-t-il des anomalies (lots ou agents) ? Si oui, lesquelles et comment les expliquer ?

Y a-t-il des différences entre agents et lots issus de secteurs d'activité différents ? Si oui, lesquelles, et comment les interpréter ? Mêmes questions en considérant les divisions administratives à la place du secteur d'activité. Y a-t-il un lien entre l'attribution d'un contrat et la distance spatiale séparant l'acheteur et le fournisseur remportant le marché ? Ou bien entre cette distance et le montant du contrat, voire une autre variable ?

Le même genre d'approche peut s'appliquer à d'autres variables, par exemple le fait qu'un agent soit une PME. Quelle est la part des PME dans la commande publique ? Par secteur ? Par type d'acheteur ? Etc.

Il est possible de calculer différents indicateurs économiques supposés révéler un fonctionnement problématique des marchés publics. De nombreux articles et rapports décrivent et discutent ce type d'indicateurs, par ex. [1]. Il peut être intéressant de calculer certains de ces indicateurs pour étudier comment ils sont distribués dans nos données. Mesurer et discuter l'association avec les variables de la FOPPA, aussi bien pour les agents que pour les lots, peut être particulièrement révélateur.

Enfin, une dernière idée est d'exploiter l'actualité. Il arrive fréquemment que des affaires de fraude dans les marchés publics soient révélées par la presse. Par exemple, récemment, l'affaire McKinsey [6] (actuellement encore en cours de traitement par la justice). Il est possible de retrouver dans la FOPPA les marchés concernés, puis de tenter d'identifier quelles caractéristiques les distinguent de la masse des marchés non-frauduleux.

Extension. La plupart des agents de la FOPPA sont identifiés par un SIRET [7], un numéro unique désignant chaque établissement commercial actif en France. Ce numéro permet de relier la FOPPA à de nombreuses autres bases de données utilisant également le SIRET. La plus importante d'entre elle est la base SIRENE [8], qui liste tous les établissements possédant un SIRET, et les décrit au moyen de nombreux attributs absents de la FOPPA. Il est donc possible d'exploiter SIRENE pour enrichir nos données de nombreuses variables supplémentaires concernant les agents économiques.

Une fois cet enrichissement effectué, il est possible de reconsidérer les questionnements précédents pour vérifier si ces données supplémentaires amènent des résultats significativement différents. Il est aussi possible d'exploiter spécifiquement les informations apportées par la base SIRENE. Par exemple, celle-ci indique si un établissement est une succursale d'une entreprise étrangère. Quelle est la proportion des marchés publics allant (indirectement) à des entreprises étrangères ? Par secteur d'activité ? Par type d'acheteur ? Etc.

Enfin, vous pouvez aussi explorer les options disponibles parmi les bases de données

5. Cf. la base SIRENE, un peu plus loin.
6. https://fr.wikipedia.org/wiki/Affaire_McKinsey
7. https://fr.wikipedia.org/wiki/Système_d'identification_du_répertoire_des_établissements
8. https://www.sirene.fr/

utilisant le SIRET, et appliquer la même méthode.

Graphes. Une autre approche est de considérer l'aspect relationnel des données de la FOPPA. On peut extraire différents types de graphes pour représenter ces relations, le plus simple étant un graphe contractuel, dans lequel les sommets représentent les agents économiques, et les arêtes les contrats entre eux. Au-delà de cette simple structure, on peut enrichir le graphe avec des attributs sur les sommets (correspondants aux informations disponibles sur les agents) et des attributs sur les arêtes (qui concernent les lots), notamment des poids. Il est aussi possible de se placer à un plus haut niveau de granularité, en agrégeant les agents (par exemple par secteur d'activité ou par zone géographique). Enfin, on peut se restreindre à des sous-ensembles des données, de manière à extraire des collections de graphes plutôt qu'un seul très grand graphe : zone géographique, période temporelle, type d'agent, secteur d'activité, etc.

Par rapport aux données tabulaires, les graphes permettent de mener des analyses supplémentaires, et offrent un support de visualisation appréciable. Pour ce qui est de l'analyse, les outils de *détection de communautés*, qui est l'équivalent pour les graphes du clustering de type k-moyennes, permettent de segmenter le graphe en sous-graphes denses. On peut alors réaliser sur les groupes obtenus le même type d'analyse que décrit précédemment. Les *mesures de centralité* sont également un outil très pertinent, permettant d'identifier les sommets les plus importants dans le graphe. On peut alors tenter de caractériser les sommets centraux, tenter d'identifier des tendances associant centralité et variables disponibles, croiser avec des catégories (secteur d'activité, zone géographique, etc.).

Pour ce qui est de la visualisation, on peut utiliser des méthodes qui vont plaquer les agents sur leurs coordonnées géographiques, afin de voir comment leurs relations sont distribuées dans l'espace physique. Mais on peut également utiliser des méthodes de spatialisation algorithmique, qui se basent uniquement sur la structure du graphe pour disposer les sommets dans un espace virtuel. Ceci permet de faire ressortir la structure des données indépendamment de la contrainte géographique.

5 Rendu

Implémentation. Vous devez fournir un script (ou un ensemble de scripts) en Python (le langage est imposé) qui, une fois lancé, effectuera l'intégralité de votre traitement à partir des fichiers originaux : préparation des données, génération des graphiques, extraction des graphes, application des algorithmes de fouille, etc. Aucune étape ne doit faire l'objet d'une intervention manuelle, de manière à pouvoir être facilement reproduit par la suite.

La manière dont ce script doit être exécuté devra être clairement expliquée à la fois dans le rapport (voir plus loin) et dans un fichier `readme.txt` à placer dans le dossier contenant le(s) script(s).

Tout ce qui peut être réalisé avec les bibliothèques utilisées en cours et TP doit l'être en priorité. Si vous avez besoin de fonctionnalités supplémentaires, vous pouvez utiliser d'autres bibliothèques que celles-ci, mais cela doit être justifié dans le rapport (et le mieux est d'en discuter oralement en séance avec l'encadrant). Tout le reste du traitement doit être implémenté dans le script lui-même.

Rapport. En plus de votre code source, vous devez rendre un rapport décrivant le travail que vous avez effectué. Vous pouvez produire ce rapport avec n'importe quel outil, tant que ce document prend la forme d'un fichier PDF, et qu'il est correctement mis en forme.

Le plan du rapport est imposé, et disponible en ligne sur Overleaf, à l'adresse suivante :

`https://www.overleaf.com/read/hbqywmgkwjgp`

Ce document n'est accessible qu'en *lecture seule*. Donc, si vous décidez d'utiliser LaTeX pour écrire votre rapport, vous devez d'abord en créer une copie avant de pouvoir l'éditer. Le rapport rendu doit être conforme aux instructions contenues dans le tutoriel suivant :

`https://www.overleaf.com/latex/templates/modele-rapport-uapv/pdbgdpzsgwrt`

Le plus simple est donc pour vous de cloner le tutoriel ci-dessus, qui est aussi un modèle Overleaf, et d'y copier-coller le code LaTeX du plan de rapport.

Notez que si vous n'êtes pas tenus d'utiliser LaTeX, en revanche, la structure du rapport est imposée, vous devez la suivre obligatoirement, en respectant les titres et la numérotation indiquée. De plus, la gestion de la bibliographie doit respecter les standards LaTeX (décrits dans le tutoriel indiqué ci-dessus).

Ressources. Vous avez le droit (et c'est même recommandé) d'utiliser n'importe quelle ressource qui pourra vous aider dans votre travail : rapports, articles, code source, pages Web, etc. La seule restriction est que vous ne pouvez pas utiliser des ressources produites par d'autres groupes de ce projet.

De plus, toute ressource doit explicitement être indiquée dans le texte de votre rapport, là où elle est pertinente. Le détail de la source bibliographique doit apparaître dans la dernière section du rapport (bibliographie), comme expliqué dans le tutoriel LaTeX.

> **Avertissement :** L'utilisation (citée ou non) d'une ressource issue d'un autre groupe, et l'utilisation non-citée ou incorrectement citée d'une ressource extérieure constituent des plagiats. En cas de plagiat, tous les groupes impliqués seront sanctionnés en conséquence. Vous trouverez plus de détail sur la notion de plagiat dans le tutoriel LaTeX cité précédemment.

Évaluation. À titre purement indicatif, le barème devrait prendre la forme suivante :

- Exploration, nettoyage, analyse descriptive de base : 8 points.
- Traitement des questionnements : 4 points.
- Extension des données, questionnements relatifs : 4 points.
- Extraction des graphes, questionnements, visualisation : 4 points.

Références

[1] J. Ferwerda, I. Deleanu et B. Unger. « Corruption in Public Procurement : Finding the Right Indicators ». In : *European Journal on Criminal Policy and Research* 23.2 (2016), p. 245-267. doi : 10.1007/s10610-016-9312-3.

[2] L. Potin, V. Labatut, R. Figueiredo, C. Largeron et P.-H. Morand. *FOPPA : a database of French Open Public Procurement Award notices*. Rapp. tech. Avignon Université, 2022. ⟨hal-03796734⟩. url : https://hal.archives-ouvertes.fr/hal-03796734.

[3] L. Potin, V. Labatut, P.-H. Morand et C. Largeron. « FOPPA : an Open Database of French Public Procurement Award Notices From 2010–2020 ». In : *Scientific Data* 10 (2023), p. 303. doi : 10.1038/s41597-023-02213-z. ⟨hal-04101350⟩.