DHLAB

1808 Venice and its real estate owners

Quentin ESTEBAN

Superviseur : Didier DUPERTUIS, Frédéric KAPLAN
Printemps 2022

Résumé :

En 1808, l'ensemble des parcelles de terrain vénitienne et leurs propriétaires sont regroupés dans un document appelé *sommarioni*. Les *sommarioni* sont un registre contenant pour chaque parcelle, des informations sur sa position, son ou ses propriétaires, sa fonction ou encore sa taille. L'objectif du projet est de regrouper les propriétaires et les parcelles dans une base de données où les liens entre entités sont explicites. Il convient de commencer par une remise en contexte historique et une présentation de ce document. Les différentes étapes du traitement des données sont : un parcours manuel du document, la création de différentes listes, notamment des termes présents, et l'utilisation des expressions régulières pour trier les propriétaires en plusieurs catégories. Puis, en se basant sur des patterns, 8124 personnes, liées à 10615 parcelles, ont pu été extraite. L'analyse des résultats obtenus à chaque étape complète la description des méthodes utilisées. Une suite logique à ce premier traitement serait d'étendre l'extraction des propriétaires aux institutions.

Introduction :

Le cadastre de venise de 1808, fournit pour chaque parcelle des informations précieuses. Ce projet consiste en un traitement de ces données par des méthodes computationnelle.

Cette contribution s'inscrit dans la création d'une base de données comprenant l'ensemble des entités présentes dans le document et l'ensemble des parcelles, où chaque propriétaire est en lien avec les parcelles qu'ils possèdent et *vice versa*. Les liens, par exemple familiaux, entre personnes doivent aussi y être préservés. Réussir à présenter les informations contenues dans les *sommarioni* sous cette forme plus granulaire serait intéressant pour faire des analyses sociologiques et historique. Cela permettrait également un accès beaucoup plus rapide, simple et intuitif aux données.

Données :

Contexte historique :

A partir de 1806, Venise est intégrée au royaume d'Italie, lui-même sous la domination de Paris. Elle perd ainsi son statut d'état au profit de celui de simple commune. L'administration française souhaite alors établir un moyen efficace de calculer et payer l'impôt. Pour cela « l'administration napoléonienne impose à la ville de Venise la mise en place d'un nouveau système de description standardisé pour rendre compte de manière objective de la forme et des fonctions du tissu urbain », comme indiqué par Di Lenoardo et al. [1]. Dès 1806, des fonctionnaires sont chargés de la collecte de données et de la création d'un cadastre.

Notre étude se base sur la version numérique d'un registre accompagnant ce cadastre, les *sommarioni, ou* états des sections. Les *sommarioni* contiennent pour chaque parcelle, des informations sur sa position, son ou ses propriétaires, sa fonction ou encore sa taille. Il convient, pour vraiment exploiter toutes les informations contenues dans ce registre, de le mettre en relation avec une carte (*mappale*, également issu du cadastre) qui permet de lier chaque entrée des *sommarioni* avec un lieu physique identifiable.

Les sommarioni :

Le point de départ pour cette étude n'est pas le document physique. Elle repose sur le travail de Isabella di Leonardo, qui a retranscrit chaque ligne du document écrit dans un fichier informatique. Cette table digitalisée permet une analyse directe avec des méthodes computationnelles.

TABLEAU 1 : EXEMPLE D'ENTREE DES SOMMARIONI

Champs	Valeurs
id	2556
sestiere	NSM
parcelNumber	1776
subParcelNumber	3
correspondanceAustrian	ASM 636 ASM 637 ASM 638
correspondanceItalian	AISM 636 AISM 637 AISM 638
toponym	SM Calle delle Balotte
houseNumber	4124
parcelOwnerText	GHENISAGHERO Enrinaldo
parcelCategoryText	Porzione di casa d'affitto
parcelCategories	["CASA"]
parcelOwnershipType	["AFFITTO"]
area	191.707
bounds	[[12.3368495,45.4360039],[12.3370319,45.4361274]]

Plus pragmatiquement maintenant, les données se trouve alors sous la forme d'un document *json*, où chaque entrée représente une parcelle, ou une sous-parcelle dans le cas où elle serait séparée en plusieurs parties. La structure d'une entrée est toujours la même et présente les différentes informations dans l'ordre suivant :

-tout d'abord, un identifiant spécifique pour chaque entrée, allant de 1 à 23427 dans l'ordre croissant, il permet de différencier chaque ligne des *sommarioni* et pouvoir les retrouver

- ensuite, des informations pour retrouver la parcelle sur la carte associée (la *mappale*, comme vu précédemment) : le *sestiere*, qui représente un des six quartiers de Venise, le numéro de la parcelle, unique pour toute la ville, le numéro de « sous-parcelle » dans le cas où la propriété d'une parcelle est divisé en plusieurs parts, ainsi que deux informations, non-présentes dans les documents originaux mais ajoutées au cours de la transcription, soit la correspondance avec les cadastres autrichien et italien. Les

deux colonnes suivantes, *toponym* et *houseNumber*, se rapprochent plus des adresses actuelles avec un nom de rue, ou de cours et un numéro de maison, unique par quartier.

- la colonne suivante est la plus importante pour ce projet, et la seul extraite pour effectuer le premier travail d'identification des personnes. La partie suivante revient plus en détail sur ses particularités.

- Viennent après la catégorie de la parcelle (sa 'fonction' en quelque sorte et le statut de la parcelle : loué ou alors occupée par son propriétaire) d'abord sous forme de texte comme écrit dans le registre manuscrit puis sous une forme standardisée avec des catégories prédéfinies.

- Enfin, les deux dernières informations sont l'aire de la parcelle et sa position sur la carte, en tant que coordonnées cette fois.

Le champ « parcelOwnerText » :

Il s'agit tout simplement du nom du propriétaire de la parcelle. C'est la catégorie avec le plus de diversité puisque toutes les autres sont soit des nombres soit des textes qui désignent une catégorie (qui sont, logiquement, moins nombreuses que le nombre de propriétaires possible différents). Pour mieux appréhender cette diversité, voici quelques exemples de textes que l'on peut retrouver dans cette colonne :

TABLEAU 2 : EXEMPLES DE TEXTES DECRIVANT LES PROPRIETAIRES

	Texte	Commentaire
(1)	GUERRA Stefano	un simple nom-prénom
(2)	GUERRA suddetto [Stefano]	suddetto « susdit »fait référence à une entrée précédente ici l'entrée (1)
(3)	PAGANI Fratelli q. Francesco	Les frères Pagani, fils de franscesco pagani
(4)	DEMANIO NAZIONALE	Domaine royale
(5)	MOROSINI Federigo ed Andrea Fratelli q. Pietro, CORNER Eredi di Nicolò, e MOCENIGO Luigi indivisi	La parcelle est partagée entre plusieurs personnes
(6)	CITTA' DI VENEZIA di propvenienza del sopresso Convento de' Padri della Certosa	La ville a récupéré la parcelle, qui avant appartenait à un couvent
(7)	BENEFICIO intitolato del primo prete ereto nella Chiesa Parochiale di S. Luca attualmente goduto dal Sacerdote Paolo GOSOLI	P. GOSOLI, prêtre de la paroisse de san Luca bénéficie de la parcelle possédée par son église.

Les propriétaires peuvent être dans les cas les plus simples des personnes seules, avec un simple nom et prénom, mais dans la majorité des cas de nombreuses précisions sont apportées. Ces particularités sont abordées dans la partie suivante concernant l'analyse du document.

Méthodes :

Parcours manuel du document et vocabulaire vénitien :

TABLEAU 3 : PRINCIPAUX MOTS DE DVOCABULAIRE VENITIEN

Mot vénitien	Signification	Exemple
Quondam, q. ou q.m	Fils de, permet d'identifier quelqu'un plus précisément que le nom de famille seul.	Foscolo Francesco quondam Leonardo
Suddetto(i)	« Susdit » * fait référence à une entrée précédente	Suddetto [Balbi Marco]
Chiesa, monastero	Eglise, monastère	Monastero di San Girolamo
S./san	Saint, diminutif de San	Chiesa di S. Vitale
Beneficio (ou benefizio) de …. Goduto dal	Un prêtre à l'usage d'un lieu mais l'église dont il dépend en reste propriétaire	Beneficio del Secondo Prete della Chiesa [di SS. Apostoli] attualmente goduto da Don Carlo Mationi
Capitolo	Chapitre religieux, assemblée de prêtre doté une personnalité juridique italienne	Capitolo della Chiesa di S. Marciliano
Regio demanio/ reale corona	Domaine/couronne royale, c'est-à-dire la propriété du royaume	-
Commune/città di venezia	La ville de Venise	-
Ministero	Ministère	MINISTERO DELLE FINANZE (des finances)
Sottoportico	Carastéristique de venise, passage couvert public traversant des édifices privés	Sottoportico di Corte Colonne
Sacerdote, prete	Prêtre	MARTINENGO Sacerdote Giovanni Battista
Primo, secondo	Premier, deuxième	Mocenigo Alvise quondam Alvise Primo
Fratello(i), sorrella(e)	Frère(s), sœur(s)	Marioni Giorgio e fratelli
Vedeva, Eredi del fu…	Veuve, héritié de feu… suivi d'une personne décédée	Torre Marianna vedova Clerle ; Eredi del fu Cassetti Niccolò
Figlio, Nipote(i)	Fils, Petit-fils	TORNIELI Giorgio e Nipoti
consorti	Consorts	DOLFIN Consorti q. Giovanni
Di provenienza (della sopressa/dal sopresso)	Lit. : d'origine, indique d'où provient la parcelle. (de la suppression de)	COMUNE DI VENEZIA di provenienza dal sopresso Monastero di S. Mauro
« di … »	-Particule pour une personne, suivant suivi d'un lieu, sorte d'équivalent du « de » français mais ne remplace pas le nom de famille	MORONI Giulio di Treviso
indivisi	La parcelle appartient à plusieurs personnes mais elle ne peut pas être divisé en sous-propriétés.	SACOGNA Giovanni, e BUJOVICH Giovanni indivisi
possessore ignoto	Propriétaire inconnu	-
sic	Rajouté lors de la digitalisation, Def : tel que cela a été dit ou écrit, sans modification, aussi étrange ou incorrect que cela paraisse	..... [sic !] di Padova

Tout d'abord, il a fallu comprendre comment était organisé le document et quelles étaient ses spécificités, plus précisément celles de la colonne des propriétaires. Cette étape peut s'apparenter à un 'défrichage'. Le parcours du document à la main a permis de voir quelle était la forme que prenait le texte décrivant les propriétaires, et aussi quels étaient les termes spécifiques du vocabulaire vénitien et /ou juridique de l'époque. Concrètement, cela a consisté en une lecture entrée par entrée d'une très grande partie du document, en notant chaque terme inconnu et en recherchant leur signification. Un entretien avec Francesca Pardini a pu répondre aux dernières interrogations. Les mots essentiels de ce début de dictionnaire pour mieux comprendre les aspects pratiques de la propriété à Venise sont regroupés dans le tableau 3.

Beaucoup d'entrée sont composées simplement du nom et prénom d'une seule personne. Cependant, ils étaient souvent accompagnés du terme « *quondam* » ou encore d'indication concernant un titre comme *sacerdote*. La parcelle peut aussi être partagée entre membres d'une même famille, notamment entre *fratelli*, *sorelle*, *nipoti*, ou *consorti*.

Le terme *suddetto* indique que, pour gagner du temps, le rédacteur du document indique que le propriétaire de l'entrée est le même qu'au-dessus. Il faut tout de même distinguer deux possibilités. La première est que *suddetto* apparait seul, cela signifie alors que le propriétaire est le même que dans l'entrée précédente. La deuxième est que *suddetto* est associé à un nom, cela signifie alors qu'il s'agit de la même personne que lorsque ce nom était apparu pour la dernière fois. Heureusement, dans la digitalisation manuelle des *sommarionni*, le propriétaire véritable de la parcelle a été ajouté à la fin de l'entrée dans des crochets.

Mais il n'y a pas que des personnes qui peuvent posséder des parcelles, un grand nombre parcelles étaient possédées par des entités publiques, c'est-à-dire la ville de venise elle-même, un ministère, ou faisaient partie du *regio demanio*. Parfois même, comme dans le cas des *sottoportico*, il n'y avait tout simplement pas de propriétaire désigné.

Suite à l'arrivée de Napoléon et à une campagne d'expropriation, de nombreux terrains sont passés sous le contrôle de la ville ou du royaume, d'où la présence de l'indication *di provenienza*.

Et enfin les différentes églises constituent le troisième grand type de propriétaire possible et celui dont les texte sont en général les plus long et complexe, car contenant beaucoup de termes techniques et de diversité dans la forme.

List des termes :

La prochaine étape a été d'établir des listes de termes (*token* en anglais) uniques afin d'avoir des chiffres concrets sur la présence des différents noms et mots. Les *list comprehension* permettent de récupérer pour chaque ligne des *sommarioni* uniquement le texte décrivant les propriétaires. Après avoir supprimé certaines ponctuations afin qu'elles ne s'agglutinent pas aux mots, le contenu des textes est

séparé à chaque espace. Cela donne des milliers de termes regroupés en une liste. A partir de cette dernière, il est possible d'obtenir tous les mots possibles et de les compter. L'analyse du dictionnaire contenant chaque mot et son nombre d'occurrence peut se retrouver dans la section résultats.

Liste des entrées :

Après avoir compté le nombre de fois où un mot apparait terme par terme, on s'intéresse contenu du champ « *parcelOwnerText* » dans son integralité. Il arrive régulièrement que plusieurs parcelles aient exactement le même contenu pour « *parcelOwnerText* ». Pour l'analyse, nous nous concentrons sur les contenus uniques. Pour la suite de ce rapport, on utilisera le terme « entrée » pour désigner un contenu unique possible pour « *parcelOwnerText* »

En utilisant le schéma simple décrit dans la partie sur le vocabulaire vénitien, le mot *sudetto* (ou *suddetti* au pluriel) est supprimé pour être remplacer par ce qu'il désigne, c'est-à-dire le propriétaire véritable.

La suppression de toutes les occurrences de *suddeti*, est complétée par celle de quelques ponctuations et espaces. Cela donne une liste d'entrées possibles dont le traitement est le même que pour les termes seuls. Le dictionnaire obtenu contient chaque texte possible décrivant les propriétaires ainsi que le nombre de fois où il apparait.

Classification des termes en catégories :

TABLEAU 4 : SET DE MOTS PAR CATEGORIES

Set	Exemples	Nombre de variantes
Membre de la famille	Fratelli, consorti, nepoti, figlio	15
Quondam	Quondam, q., q.m, q.m.	4
Titre	Sacerdote, primo, secondo	4
Prénoms	Giovanni, Carlo, Luigi	331
Nom de famille	Agostini, Contarini, Fontanella	1725

Pour les sets de titre, de membre de la famille et de façon d'écrire *quondam*, un simple inventaire manuel est suffisant

Pour établir les listes de prénoms et de noms, deux principales méthodes ont été utilisées. Premièrement, des mots parmi ceux revenant le plus souvent dans les deux premières listes crées (celles des mots et celles des entrée) ont été ajoutés, en considérant que ceux en majuscule étaient des noms de famille et ceux en minuscule des prénoms. Ensuite une observation des entrées non-classées après chaque itération de création des personnes (dans l'étapes d'extraction des propriétaires) permet de compléter au fur et à mesure les sets. Bien sûr, avant l'ajout d'un nom, une rapide vérification de sa cohérence dans les *sommarioni* et sur internet (registre des

noms courant italien *etc*) est nécessaire. Cette méthode a permis d'éviter de considérer à tort des noms de famille comme des prénoms parce qu'ils étaient écrits en minuscule par exemple.

Liste des bigrammes :

Une autre manière intéressante de comprendre les données est d'étudier les bi-grammes. Les N-gramme sont une suite de n mots, ou *token* plus généralement, qui se suivent, par exemple des nom-prénom ou d'autres enchainements plus ou moins courant comme « *di venezia* » ou « *e fratelli* ». Une bibliothèque utile pour les extraire est *nltk*. Le texte est d'abord séparé en *token*, qui correspondent aux mots. *Nltk* permet ensuite de créer une liste des *bigrammes*. Qu'encore une fois il convient de transformer en un dictionnaire contenant les couples de mots possibles et leur nombre d'apparitions.

Tri des propriétaires en catégories :

TABLEAU 5 : DIFFERENTS TYPE DE PROPRIETAIRES POSSIBLES

Catégories	***Descriptions***
Institution :	
Venezia	La ville elle-même
Demanio	Domaine royal
Chiesa	Institution religieuse
Congregazione	Congrégations religieuses
Une personne :	
Nom Seul	Une seule personne est référencée
Nom avec quondam	Une personne seule et son quondam
Nom avec famille	Une famille qui partage le même nom
Autres :	
Sic	Champ avec potentiellement des erreurs
irrelevant	Champ vide ou équivalent
Autres	Ce qui ne rentre pas dans les catégories précédentes

La prochaine étape consiste à voir quel type d'entité possède le plus grand nombre de parcelles à partir du dictionnaire des entrées possibles crée plus tôt. En prenant en compte les informations récoltées et l'inspection manuelle et computationnelle des *sommarioni*, avec des expressions régulières notamment, la classification suivante, visible dans le tableau 5, est établi et va être utilisé pour la suite :

- Venezia : les parcelles appartenants à la ville elle-même. Cela se traduit par trouver les parcelles qui contenait *città* ou *commune di venezia*, à première vue, comme il s'agit du même propriétaire à chaque fois, on s'attend à avoir une diversité des entrées très faible avec uniquement une poignée d'entrée. Mais ce n'est pas le cas puisque beaucoup d'entrée fournissent des informations

complémentaires sur le passé de la parcelle, souvent sous la forme d'un ajout commençant par *provenienza* et sont donc considérées comme différentes.

- Demanio (regio) : littéralement domaine royale. Plus précisément c'est ce qui est la propriété du royaume. Cela concerne les entrées avec « *regio demanio »* mais aussi « *ministero »* (des finances, de la guerre, etc) ou « *reale corona »*. Le cas du demanio est similaire au précèdent, il y a beaucoup de précision sur le modèle « *di provenienza* »
- Chiesa : Le cas le plus complexe. Les propriétés religieuses sont désignées de manière très diverse, et même celle d'une même église peuvent être dénommées de plusieurs façons. Le cas des congregazzione est particulier. Malgré leur présence modeste et bien qu'elles puissent se rapprocher des églises, elles sont triées séparément.

A présent, la grande majorité des parcelles restantes désignent des personnes physiques et non plus des entités juridiques. Les regrouper toutes ensemble et en faire une très grande catégorie ne semble pas pertinent. Cela ne donnerait pas beaucoup d'informations et une délimitation plus fine permettra pour la suite de plus facilement extraire les noms, prénoms, titres, *etc* des propriétaires. Les personnes sont donc réparties en trois grands groupes :

- Le plus simple, nomSeul : une personne seule, concrètement cela se traduit par un mot (typiquement le nom de famille) suivi ou non par un ou deux autres mots, auquel j'ai ajouté la possibilité de commencer par « *eredi del fu* » et de contenir la particule « di ». L'idée est de récupérer toutes les entrées simples
- Viennent après les entrées de personnes seules mais qui contiennent le terme « *quondam* », ce qui ajoute une référence à une nouvelle personne, peut-être décédée mais qu'il sera pertinent d'ajouter à notre analyse
- Enfin, on a aussi des noms qui ne sont pas seul mais avec des membres de leur famille.

Après les avoir classés en utilisant ces catégories principales, il restait encore quelques entrées. Elles sont séparées en :

- Sic. Elle concerne les entrées concernant le mot « sic » ou une longue suite de point, ce qui indique que la fiabilité du texte peut être remis en question ou qu'il a été tronqué car illisible.
- Irrelevant, ou inutile car elle contient les entrée vides ou consistant en un seul tiret, donc impossible à relier à une entité précise. Cependant, le fait de savoir que le texte est vide peut aussi nous indiquer que la parcelles n'est la propriété de personnes dans certain cas (ex : *sottoportico*)
- Autres : tout ce qui n'a pas été classé jusqu'à présent est regroupé ici. Il est normal de s'attendre a priori à un groupe incohérent. Cependant ce sont souvent des textes contenants simplement plusieurs personnes à la suite. Il pourrait presque être renommer en NomPlusieurs, en référence à nomSeul.

L'utilisation des expressions régulières permet de vérifier si un texte rentre dans un format prédéfini. Dans ce cas, à la liste correspondante est ajouté le tuple (texte, nombre de parcelles). Dans le cas contraire, le texte est comparé à la prochaine

catégorie. L'ordre des tests est le même que dans la présentation ci-dessus. Ainsi les catégories sont *de facto* mutuellement exclusives.

Une itération sur l'ensemble des entrées résulte en la création de plusieurs listes, une par catégorie. Ces dernières fournissent des statistiques intéressantes : le nombre d'entrée différentes et le nombre de parcelles total possédée dans chaque catégorie. Ces deux statistiques permettent de créer un graphique présenté dans la section résultats (Figure 1) pour de se rendre compte de l'importance relative des différents types de propriétaires.

Attention, le nombre d'entrée n'est pas forcément le nombre d'entités différentes puisqu'on compte ici les textes strictement identiques. Par exemple la *commune di Venezia* est appelée de plusieurs manière et compte pour plusieurs, tandis que deux homonymes compteront pour une seule entrée. Cependant il est probable que pour les entrées concernant un nombre limité de personnes, c'est-à-dire un individu et sa famille, cela correspond à peu près à la réalité.

Extraction des propriétaires :

Cette façon de comptabiliser les parcelles fournit donc des informations quantitatives sur les parcelles concernées mais pas qualitatives. Il est possible de savoir combien chacun possède mais pas ce qu'il possède exactement. Cela est plus simple et plus économe en mémoire de stocker le tuple (texte, nombre de parcelles), mais retrouvez ensuite les propriétés liées à une entrée est difficile.

Pour remédier à ce problème, le simple nombre de parcelles du dictionnaire précédent est remplacé. A la place, la liste des parcelles possédée par chaque entrée est conservé sous la forme d'une liste d'*id* (les mêmes que dans le *json* original). Comme le nombre de parcelles n'était pas utilisé pendant le tri en catégories, la même méthode est employée.

Les données sont alors prêtes pour la partie finale de ce projet, qui est de crée des entités de propriétaires et de les relier à leurs possessions. On ignore les « institutions » et « autres » du tableau 5 pour se concentrer sur les personnes.

Les personnes, ou familles, sont représentées par des objets. Ses attributs sont les suivants :

TABLEAU 6 : ATTRIBUTS D'UNE PERSONNE

Attributs	*Descriptions*
Noms :	
Nom de famille 1	Nom de famille, seul attribut obligatoire
Nom de famille 2	En cas de double nom
Prénom 1	Le prénom
Prénom 2	Le second prénom éventuel
Di	Mot qui vient après un « di »
Membre de la famille	Relation qui unit le groupe
Titre	Ce qui définit autre que le prénom
Booléens :	
isQuondam	Si il est le quondam et non le propriétaire
isMorto	Si il est décédé
Parcelles :	
seul	Possédée seul
Avec quondam	Possédée et le quondam est précisé
Avec famille	Possédée avec un ou des membre de sa famille
Avec Autres	Possédée avec des inconnu
Indivisi	Indivisible
Liens :	
lien	Personnes liées et leur lien respectifs

- Les attributs de la partie parcelles du tableau sont des liste d'id
- L'attribut lien est une liste de tuples (personne, lien)

Les textes de propriétaires ont des formats récurrents. Cela peut-être, par exemple, Nom-prénom (comme dans Tableau 2 (1)), Nom-prénom-*quondam*-prénom (Tableau 2 (3)), ou nom-prenom-membre. Pour extraire les personnes, trois étapes sont nécessaires : 1) la séparation du texte par mots 2) une comparaison de l'entrée avec un des formats connus. 3) Si un format est reconnu, la création d'une nouvelle personne, ou plusieurs dans le cas de *quondam*, *fratelli* etc.

Chaque entrée classée doit être de bonne qualité. Pour s'en assurer, la détection des formats (étape 2) se fait en vérifiant l'appartenance des mots à un set pour chaque attribut de « Noms » de la classe Personne (à l'exception de l'attribut Di, où la présence de mot « di » suffit à l'identification). Par exemple, une entrée est considérée comme de type nom-prenom : -si elle contient deux mots -que le premier mot appartient au set des nom de famille – que le second mot appartient au set des prénoms.

Résultat :

Liste des termes :

Parmi les mots qui reviennent le plus souvent, on retrouve :

- des prénoms comme *Giovanni* (3903 fois), *Antonio* (2687), *Pietro* (2117) ou *Francesco* (2061) et des noms, mais avec beaucoup moins d'occurrences, comme MOROSINI (382), CONTARINI (340)

-des indications de relations comme *fratelli* (1513) ou *vedova* (579)

- *quondam* (3761) ainsi que ses variantes q. (3532), q.m. (3252), q.m (3118).

- des articles comme *di* (3664), *e* (3300), *del* (962)

- des mots spécifiques au document comme *indivisi* (1144) ou *suddetto* (1567)

- des termes qui se rapportent à la ville et au royaume ; *VENEZIA* (1088) *Venezia*(519), *DEMANIO* (221)

- et enfin des mots associés au religieux ; *S.* (1557), *monastero* (323), *CAPITOLO* (200), *sacerdote* (194)

Cependant, on peut remarquer que très peu de termes reviennent plus de 100 fois (environ 150 sur 6740) et encore moins plus de 1000 fois (23 sur 6470) la grande majorité des mots sont présents moins de 10 fois, dont un tiers qu'une seule fois, et un autre tiers que deux fois. Parmi les mots présents qu'une seule fois, il y a des noms de famille, souvent, mais aussi occasionnellement des erreurs de numérisation, ou en tout cas probablement car il s'agit de variations orthographiques 'exotiques' de mots présents beaucoup plus souvent. La principale information de la liste des termes est que les *sommarioni* ont une grande diversité théorique de mots mais dans la pratique, une poignée seulement occupent une grande partie des textes.

Liste de bigrammes :

Les bigrammes le plus fréquents sont surtout en lien avec des institutions : religieuses avec « *di S.* » (1198 fois), « *della chiesa* »(341), « *monastero di* » (212) , « *chiesa di* » (164), ou communale/étatique avec « *DI VENEZIA* »(1083), « *di Venezia* »(515), « *di provenienza* »(679), « *Venezia di* »(361)

Hors les prénoms composés comme « *Giovanni Battista* » (1033), les bigrammes avec beaucoup d'occurrences pouvant être associé à des personnes sont « *possessori indivisi* »(436), « *e fratelli/Fratelli* », (374+375) ou « *eredi del* »(248).

En général, ces bigrammes soit contiennent *e, di* ou leurs dérivés, soit sont issu d'une phrase ou d'une formulation type qu'on peut alors retrouver en observant ceux qui ont une fréquence similaire. Par exemple évidemment « *commune di venezia* », mais aussi « *di provenienza della sopressa* » ou « *e per esso il* ».

Liste des entrées :

Le fait que l'entrée possible soit associée à leur nombre d'occurrences permet de les trier en fonction de ce dernier. Une première observation est que les cinq premiers textes les plus fréquents désignent tous la ville de Venise (pour un total de 683 fois). Il y a ensuite « *possessore ignoto* » (48 fois). Les suivants sont un « *capitolo* » (36) une « *congregazione* » (33) et « *MARUZZI Costatino* » (29). Une conjecture raisonnable est qu'avec 29 parcelles, cet homme est le plus grand propriétaire 'privé' de la ville (en excluant la possibilité d'une même personne désignée différemment selon les entrées, ce qui est possible). Aussi, comme attendu, les fréquences sont très faible comparées à celles des termes, ce qui est normal puisque beaucoup plus restrictif. La majorité des entrée possibles n'apparaissent qu'une seul fois (environ 75 %).

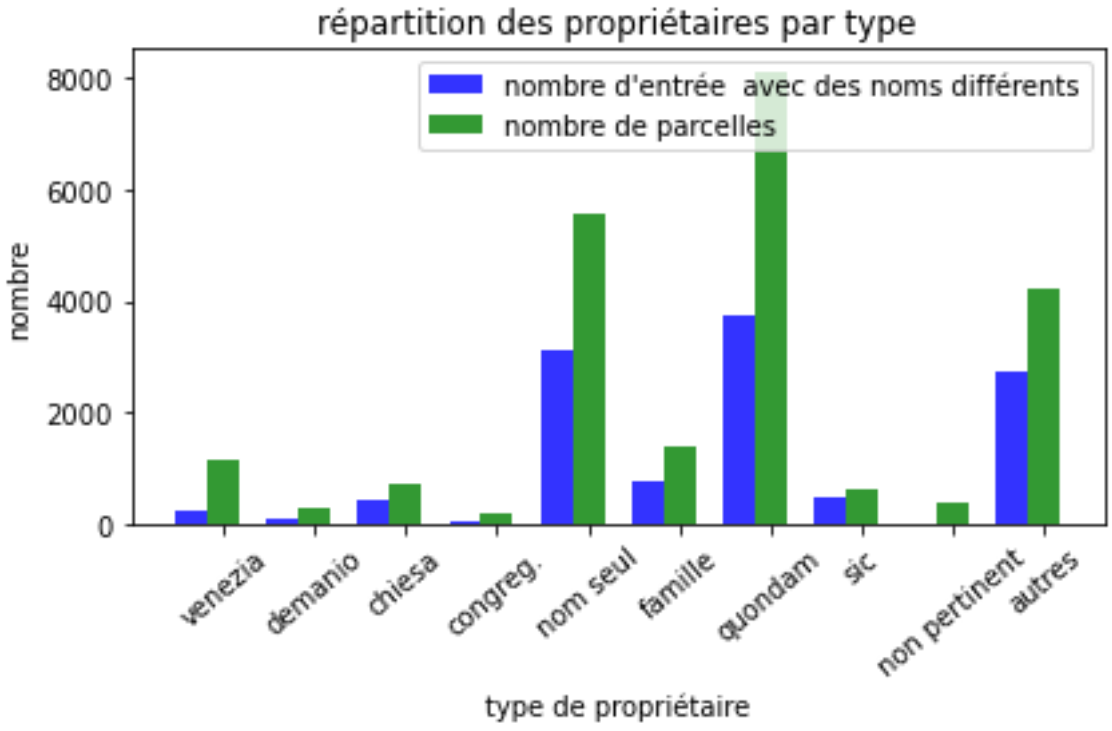

FIGURE 1 REPARTITION DES PROPRIETAIRES PAR TYPE

Ce graphique confirme l'observation dans l'analyse de la liste des entrées concernant la ville de Venise. Un petit nombre de façon de l'écrire par rapport aux parcelles. C'est le même cas pour *demanio* en moins prononcé. Tout le contraire de la catégorie des *chiesa*, où la diversité d'appellation est plus visible que nulle part ailleurs, ce qui explique aussi une plus grande difficulté de classification. En ce qui concerne les noms, chaque entrée possède en moyenne entre 1.5 et 2 parcelles ; peut-être environ le nombre de terrains moyen d'un propriétaire vénitien ?

Les catégories des noms regroupent les deux tiers des parcelles, soit la majorité. Les institutions quant à elles, possèdent moins d'une parcelle sur 10.

Classification des personnes :

TABLEAU 7 : ETAT DE LA CLASSIFICATION

	Nom seul	Nom - Quondam	Nom avec Famille	Total
Total entrée	3104	3713	745	7562
Entrées non classées	528	1073	189	1790
Entrées classées	2576	2640	556	5772
Pourcentage classé	82.9%	71.1%	74.6%	76.3%
Personnes créées	**2576**	**4509**	**1039**	**8124**
Total parcelles	5153	7293	1157	13603
Parcelles non classées	811	1890	287	2988
Parcelles classées	4342	5403	870	10615
Pourcentage classé	**84.2%**	**74%**	**75.1%**	**78%**

Le nombre de personnes crées est plus élevé que le nombre d'entrées pour les deux dernières colonnes car les texte font parfois référence à plusieurs personnes.

Le taux de classification des parcelles est plus élevé que celui des entrées. Cela s'explique en partie par le fait que les noms de famille ajoutés au set des noms ont d'abord été ceux concernant le plus de parcelles.

Au total, pour les entrées concernant au maximum une famille, plus de 76% des entrées, qui en tout représentent 78% des parcelles, ont été classées et liées à un total de 8124 personnes.

Obstacles rencontrés :

Les termes spécifiques au langage juridique du cadastre, combiné au Vénitien ancien a parfois été dur à comprendre. Il y a eu méprise sur le sens exact du terme pourtant essentiel de *quondam*, qui a mené a changer d'implémentation.

Les accents italiens ont causé quelque souci d'encodage, notamment pour lire les *json* ou y rechercher des termes.

Utiliser des sets de nom de famille (dans l'étape d'extraction des propriétaires) est efficaces mais passé les noms les plus courant, l'ajout progressif est laborieux et de moins en moins efficaces en termes de nouvelles entrées classés

Conclusion :

Du traitement des données nécessaire avant la création des personnes sont ressortis beaucoup d'information. Des tendances sur les propriétés foncières ont pu être conjecturée, par exemple un propriétaire privé possèderait en moyenne entre 1.5 et deux parcelles.

Quant à l'extraction des données, 76% des entrées, dans les catégories appartenant à « Une Personnes » de la table 5 ont été classé. Cela concerne 78% des parcelles de la catégorie. Il en résulte l'ajout à la base de données de 8124 Personnes et 10615 parcelles.

La création de la base de données parait donc possible et souhaitable :

- pour confirmer et affiner les conjectures
- également car elle serait beaucoup plus simple à manipuler que les textes, qui ne sont standardisés que pour l'œil humain. Il serait possible par exemple, de choisir toutes les personnes ayant le même nom de famille et de les colorer sur la *mappale*.

Cependant, plusieurs limites ont pu être entrevu :

- La classification en une seule catégorie par entrée ne permet pas de rendre compte des cas, certes rare, où une parcelle est possédée à moitié par une institutions et à moitié par une personne.
- A terme, il serait désirable de pouvoir identifier si un même individu possède plusieurs parcelles. Or même si deux parcelles ont le même texte, il est impossible de savoir si ce ne sont pas simplement des homonymes.
- A l'inverse, il est possible qu'une personne soit présente dans le document dans des formats différents : par exemple une fois seule et une fois avec un membre de sa famille. Dans ce cas, plusieurs personnes seront créées à partir d'un seul propriétaire réel.

Travail futur :

Un traitement ultérieur des personnes pourrait résoudre le problème précédant. Il serait possible de comparer les attributs des personnes crées pour les fusionner, par exemple s'ils ont le même prénom, nom et qu'ils ont le même père (« quondam »). Cependant, dans l'absolu, il sera toujours impossible de savoir s'il s'agit de la même personne.

Un autre ajout à ce travail pourrait être de reconnaitre, dans les entrées des autres catégories, notamment « Nom Plusieurs », les personnes déjà présente dans la base de données.

Aussi, seuls des ébauches de classifications, non présentées ici, ont été implémentées concernant les institutions et les entrées dite spéciales (compagnie des marchands, cimetière, *sottoportico*…). C'est la continuité logique directe de ce travail et l'axe principal d'amélioration du modèle.

Bibliographie :

[1] Isabella di Lenardo, Raphaël Barman, Federica Pardini et Frédéric Kaplan, « Une approche computationnelle du cadastre napoléonien de Venise », *Humanités numériques* [En ligne], 3 | 2021, mis en ligne le 01 mai 2021, consulté le 12 mai 2021. URL : http://journals.openedition.org/revuehn/1786 ; DOI : https://doi.org/10.4000/revuehn.1786