# De la extracción de datos a la representación geográfica. Una propuesta de utilización de documentos obtenidos mediante transcripción automática

Rosa Congost, Ricard Garcia Orallo, Jordi Regincós, Enric Saguer y Lluís Serrano
(Centre de Recerca d'Història Rural, Universitat de Girona)

El Centre de Recerca d'Història Rural de la Universitat de Girona está trabajando desde 2016 en un proyecto de largo alcance de transcripción automática de los libros del Oficio de Hipotecas de Girona entre 1768 y 1805. El proyecto, que conlleva un importante desafío tecnológico, se ha podido llevar a cabo gracias a la colaboración con el *Pattern Recognition and Human Language Technology Research Center* (PRHLT) de la Universitat Politècnica de València, un equipo especializado en las técnicas de reconocimiento de texto manuscrito. Esta tecnología aún se halla en estadio de desarrollo, a diferencia del reconocimiento de texto impreso (OCR), que ya ha alcanzado un nivel de madurez considerable. Hasta el momento se ha contado con el apoyo del programa de Humanidades digitales de la Fundación BBVA y del Ministerio de Ciencia, Investigación y Universidades.[1]

La fuente objeto del trabajo son los libros del Oficio de Hipotecas correspondientes al distrito hipotecario de Girona. El Oficio de Hipotecas, creado mediante una Real Pragmática de 31 de enero de 1768, tuvo como objetivo registrar y publicitar todas las escrituras notariales que estuvieran gravadas con cargas, fianzas o hipotecas. La interpretación de este mandato dio lugar a distintas prácticas en lo que se refiere a su forma y contenido. En el caso de Cataluña, si bien la aplicación de la norma también ofrece resultados territoriales dispares, parece que indujo al registro sistemático de muchas tipologías documentales. Tanto el celo expresado por los colegios notariales y la demanda de seguridad registral como las resoluciones que se dictaron a petición de la Real Audiencia de Cataluña, impulsaron una ampliación generosa del registro hacia todos aquellos contratos que contuvieran hipoteca general para alguna de las partes contratantes, o algún tipo de fideicomiso. El registro constituye, desde esta perspectiva,

[1] El se ha realizado dentro del proyecto PGC2018-096350-B-I00, financiado por el Ministerio de Ciencia, Innovación y Universidades, y el Fondo Europeo de Desarrollo Regional de la Unión Europea. En fases previas también ha contado con una ayuda del programa de Humanidades Digitales de la Fundación BBVA (2017-2018)

una valiosa fuente para el estudio tanto del mercado del crédito y de la tierra como de muchos otros aspectos de la vida social.

El conjunto documental en el cual se ha focalizado el proyecto está integrado por un total de 75 gruesos libros, cuya digitalización ha dado lugar a 104.718 imágenes. Actualmente se dispone de la transcripción completa de los 28 primeros libros, correspondientes a los años 1768-1780 y al 50% de las imágenes. El producto resultante es un conjunto de archivos de texto plano en formato XML, además de otro conjunto de archivos XML que contiene, junto con la transcripción línea a línea, la geometría del texto en cada imagen.

La metodología y las herramientas del proceso de transcripción automática, así como sus primeros resultados, ya fueron expuestos en una comunicación presentada al congreso de la SEHA de 2018 (Bosh, Congost, Quirós, Saguer y Vidal, 2018). Por ello, aunque ahora dispongamos de más información y experiencia sobre los resultados de la transcripción, nos permitiremos centrar nuestra atención en la fase siguiente del proceso de trabajo y análisis. El tema que vamos a desarrollar se refiere a cómo nos planteamos tratar la gran masa de documentos resultantes del proceso de transcripción. En términos generales, la gestión y utilización de gran cantidad de datos que está aportando la digitalización de las fuentes documentales genera nuevos retos al historiador, quien se ve urgido a desarrollar procedimientos y estrategias para utilizar esta información. En nuestro caso particular, a los retos generales se añade otra dificultad: el manejo de textos obtenidos mecánicamente que, si bien son generalmente comprensibles, inevitablemente contienen errores de transcripción.

De acuerdo con los objetivos de la sesión, enfocaremos la comunicación hacia una cuestión más concreta que nos permita enlazar la cuestión de transcripción automática y su explotación con las técnicas de representación cartográfica. Nos centraremos, pues, en cómo extraer masivamente datos geográficos contenidos en la documentación transcrita con el objetivo de localizar territorialmente los hechos registrados en el Oficio de Hipotecas y de realizar un análisis espacial de algunas variables vinculadas al mercado del crédito. Los resultados presentados, como puede deducirse de lo ya expuesto, son necesariamente provisionales y sólo pretenden mostrar una línea de trabajo que esperamos que de pronto mayores frutos.

## De la transcripción automática a la base de datos: preparación de los archivos

El proceso de reconocimiento de texto toma una colección de imágenes digitales discretas correspondientes, cada una, a una página de libro y devuelve, como resultado, una colección también discreta de archivos XML que contienen el texto y su geometría. Como cabe suponer, pocas de estas imágenes corresponden al registro de un único

documento. En algunas ocasiones una imagen contiene más de un asiento; y en otras, el asiento completo se encuentra en varias imágenes. Una imagen digital contiene, pues, las más de las veces, fragmentos de varios documentos. La primera tarea que debe resolverse consiste en reunir los archivos individuales de un mismo libro y (re)construir las unidades documentales que contiene, ya que estas serán nuestras unidades básicas de procesado de los datos. Para ello se utiliza información geométrica y también el etiquetado automático que genera el mismo programa de transcripción, y que identifica –entre otros elementos– la tipología documental que figura de manera sistemática en los márgenes del libro. La reconstrucción documental también consiste en reubicar como notas al pie del documento correspondiente todas aquellas otras notas marginales que contienen enmiendas o información sobre trámites realizados con aquel asiento.

La reconstrucción documental es el primer paso para integrar los documentos en una base de datos y, después, extraer aquella información susceptible de ser organizada de forma serial. Para alcanzar este objetivo, los documentos son descompuestos palabra a palabra (tokenización) y el resultado es sometido a un proceso de depuración y estandarización. Actualmente, los 28 libros transcritos han generado una tabla con algo más de 21 millones de tokens o palabras. Las funciones que se realizan a lo largo de este proceso de preparación de los documentos son, básicamente, las siguientes:

- recomposición de las particiones de palabras entre líneas
- separación del texto respecto de las etiquetas que el propio programa de transcripción ha generado. Algunas de estas etiquetas marcan abreviaturas y las despliegan, y otras identifican algunos tipos concretos de datos (topónimos, antropónimos, oficios y tipologías documentales).
- revisión de los términos que deben ser etiquetados como topónimo o como oficio.
- reconstrucción de los términos compuestos, especialmente nombres propios y topónimos.
- conversión de las fechas absolutas y relativas a un formato estándar de aaaa/mm/dd
- conversión de las expresiones monetarias en valores numéricos

El paso siguiente consiste en identificar y desambiguar aquellos elementos que puedan tener valor analítico. No es suficiente saber que una palabra es un topónimo, sino que debemos poder detectar cual es la función de dicho topónimo en el texto: desde el lugar de registro del documento hasta la residencia del vendedor en operaciones de compraventa. Por ello, distinguimos conceptualmente entre *ítem de información* (que nos sirve para determinar si una palabra o una secuencia es un antropónimo, una etiqueta sociolaboral, una fecha,...) y *dato* (que precisa cual es el sentido o la función de

dicho ítem en el texto). Hasta el momento se han desarrollado algoritmos para distinguir e identificar datos de carácter general, que aparecen en todos los asientos y no sólo los de un tipo documental específico. Concretamente, por el momento podemos identificar automáticamente la fecha de la escritura notarial, la fecha de registro en el oficio de hipotecas, el nombre del notario, la categoría sociolaboral con la que aparece asociado, la localidad de la notaría y la función del notario (especialmente cuando se trata de una regencia).

**Identificación y desambiguación de topónimos: el lugar de la notaría**

Dado que nuestro objetivo es poner el foco sobre los elementos geográficos, expondremos las características y dificultades del proceso de extracción del lugar de la notaría. A diferencia de otros topónimos, su identificación es relativamente simple porque siempre aparece en las primeras líneas de cada nuevo asiento. El arranque, ya se trate de una operación de compraventa, un testamento o unos capítulos matrimoniales, se repite sistemáticamente y es del siguiente tenor:

> *Arriendo | En nueve Febrero mil Setecientos y Setenta se ha presentado en el Oficio de Hipotecas de esta ciudad de Gerona una escritura otorgada ante Juan Bautista Morell y Milsocos Notario de la villa de Figueras en catorze Enero proximo passado...*

Una vez realizadas las operaciones preparatorias que hemos expuesto, el texto del documento puede ser reconstruido como sigue:

> *Arriendo En $dataAbs[1770/02/09] se ha presentado en el Oficio de Hipotecas de esta ciudad de **$llocAbs[Girona]** una escritura otorgada ante $antroponim[juan bautista morell y milsocos] $ofici[notario] de la villa de **$llocAbs[Figueres]** en $dataRel[1770/01/14]...*

La primera referencia toponímica siempre corresponde a la localización del Oficio de Hipotecas –que, en nuestro caso, mientras no ampliemos el alcance del trabajo, va a ser Girona–, y la segunda, habitualmente, a la notaría. Ambos topónimos, como se observa en el ejemplo, han sido correctamente identificados con la etiqueta **$llocAbs**, que denota un topónimo literal o absoluto. Sin embargo, con mucha frecuencia las referencias son relativas. En vez de indicar el topónimo, cuando este ya ha sido mencionado anteriormente, aparecen expresiones como ‘en dicha ciudad’ o ‘en la misma ciudad’.

> *Establecimiento En $dataAbs[1768/06/02] se ha pntado en el Oficio de Hipotecas de esta Ciudad de $llocAbs[Girona] una Escritura otorgada ante $antroponim[francisco casanoves y garriga] $ofici[notario] **de dicha Ciudad**, en $dataRel[1768/05/28]...*

Para estos casos, se está desarrollando un algoritmo de sustitución que lea el topónimo previo y reemplace la referencia relativa por su valor. El fragmento anterior, una vez reconstruido, quedaría así:

> *Establecimiento En $dataAbs[1768/06/02] se ha pntado en el Oficio de Hipotecas de esta Ciudad de $llocAbs[Girona] una Escritura otorgada ante $antroponim[francisco casanoves y garriga] $ofici[notario] de* ***$llocRel[Girona],*** *en $dataRel[1768/05/28]...*

Otra dificultad aparece cuando la notaría no se encuentra en la segunda posición del orden de topónimos, como en el caso de algunas regencias donde se indica antes el lugar de residencia del notario regente y este no coincide con el lugar de la notaría:

> *Venta En $dataAbs[1768/08/25] ve se ha presentado en el Oficio de Hipotecas de esta Ciudad de $llocAbs[Girona] una escritura otorgada ante $antroponim[pedro puig]* ***$ofici[notario] de la Villa de $llocAbs[Figueres] regentante la Notaria del Castillo de $llocAbs[Siurana]***

La existencia de palabras clave, en este caso 'regentante' o 'notaria', nos pueden permitir ir refinando los algoritmos para resolver los casos más complejos, como el que acabamos de ver, aunque siempre quedará una porción sin resolver, que se sumará a algunos errores de transcripción que impiden la detección de la notaría. A pesar de ello, los 28 libros transcritos hasta el momento arrojan un resultado muy satisfactorio. El procedimiento automático ha detectado las localizaciones notariales en el 84,1% de asientos, y mediante un procedimiento posterior de revisión asistida se ha conseguido aumentar el resultado hasta el 99,3%.

Esto nos permite obtener con relativa facilidad una primera visualización cartográfica de algo tan simple como la actividad de cada notaría durante el periodo examinado. Para ello es necesario realizar dos operaciones previas. La primera consiste en homogeneizar y estandarizar la denominación toponímica, subsanando tanto las ambigüedades del propio texto como los errores de transcripción. La segunda es asignar a cada valor toponímico una georreferencia específica. Para ello se ha utilizado el *Nomenclàtor oficial de toponímia major de Catalunya* (edición 2009), que contiene las coordenadas UTM.

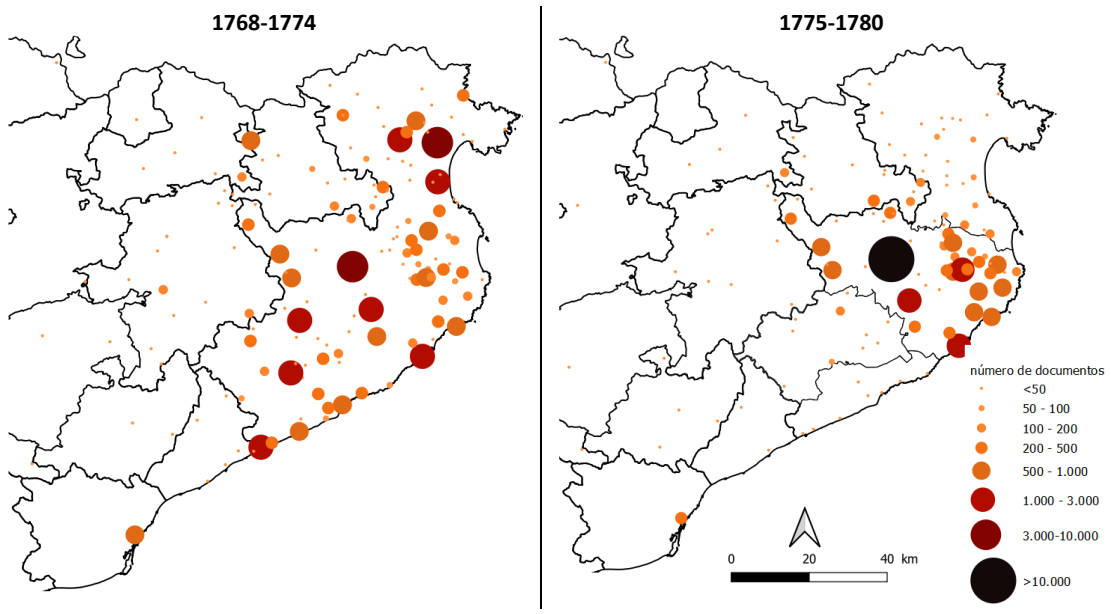
Antes de comentar los resultados, debe advertirse que, a los pocos años de su creación, el distrito hipotecario de Girona empezó a desgajarse, dando lugar a nuevos distritos y reduciendo el territorio inicial del distrito gerundense. Los detalles de este proceso fueron expuesto en la comunicación presentada el 2018 y no vamos a repetirlos, pero son la razón por la cual hemos generado una doble serie de representaciones cartográficas, distinguiendo entre dos periodos de amplitud muy similar (1768-1774 y 1775-1780).

De la figura 1 pueden sacarse dos conclusiones de interés. La primera se refiere a la vinculación entre notaría y territorio. Como es sabido, no existía obligación de escriturar los actos referidos a una finca o a un patrimonio en una notaría específica por razones de cercanía o de distribución territorial. Los actuantes, en principio, podían registrar sus decisiones o acuerdos ante el notario que más les conviniera, sin importar la distancia respecto de los bienes implicados. Los libros del Oficio de Hipotecas nos ofrecen la

posibilidad de cuantificar y valorar el alcance de esta práctica dado que la Pragmática de 1768 establecía que se debía tomar razón de la escritura en los pueblos –o sea, cabezas de distrito hipotecario- "en que estuvieron situadas las Hypothecas (...); y si las Hypothecas estuvieren situadas en distintos Pueblos se anotará en cada una de las que les correspondan" (art. 1).

Como se observa en los mapas de la figura 1, la mayor parte de la contratación se realizaba dentro del mismo distrito hipotecario o en las zonas limítrofes a él. El ámbito de actuación de los notarios, pues, solía tener un radio limitado. Sólo el 4,3% de los contratos que se registraron en los libros del Oficio de Hipotecas procedían de notarías externas al distrito, entre las cuales tenía un peso relevante las de la ciudad de Barcelona. Este porcentaje aún disminuye con el tiempo, cuando lo medimos por etapas (1768-74: 4,6%; 1775-80: 3,9%), lo cual estaría indicando que el radio de influencia de cada notaría era bastante reducido y que la fragmentación territorial, al contrario de lo que podríamos haber esperado, no lo afectó.

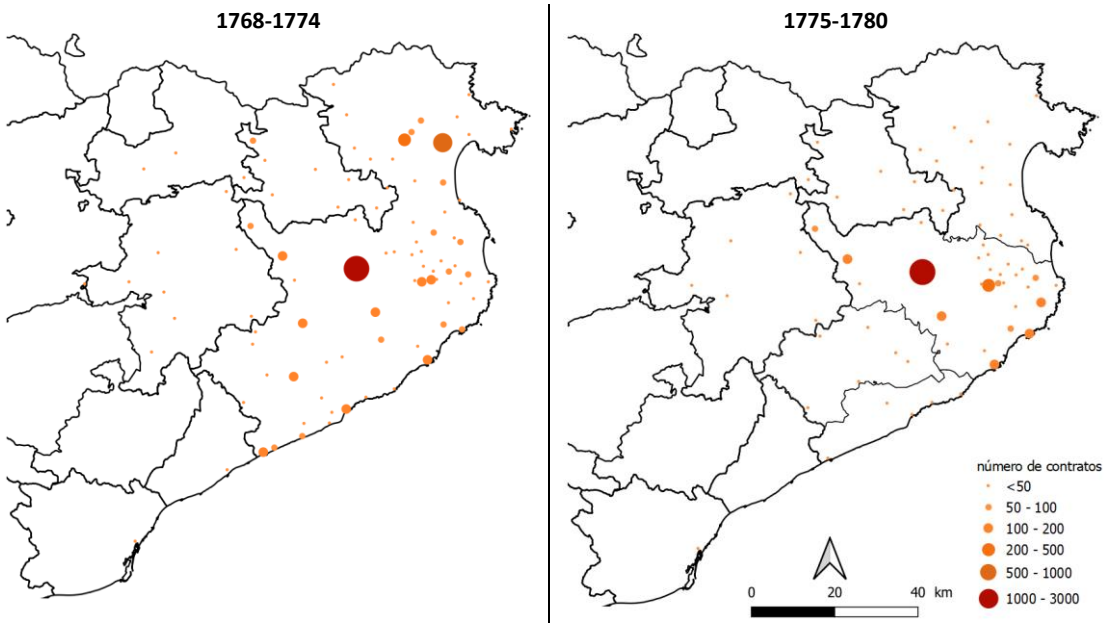
**Figura 1: Actividad notarial**




La segunda observación que nos brinda la figura 1 se refiere a la concentración de la actividad notarial. A pesar de que la ciudad de Girona acumulaba hasta once notarías simultáneas, el nivel de concentración de la actividad notarial era moderado. Existía una tupida red de notarías de mediana y pequeña dimensión con un nivel de actividad notable, a juzgar por los datos del registro de hipotecas. Para el periodo 1768-74, cuando el distrito hipotecario incorporaba algunos centros importantes, como Castelló d'Empúries, Arenys de Mar, Sant Hilari Sacalm o Figueres, los notarios de Girona sólo

tramitaron el 12,5% de los contratos que se presentaron al registro. En el segundo periodo, con un territorio bastante menor y sin los núcleos mencionados, el peso de los notarios de Girona aumentó sustancialmente, hasta el 41,9%. El aumento no fue sólo relativo, sino también absoluto. Sin que podamos explicar la razón, la cantidad de contratos de estos notarios se multiplicó por 2,5 para un intervalo temporal similar (5 años) al primer periodo (5,5 años). Aun así, continuó persistiendo la red de notarías medianas y pequeñas, y, aunque menos, también en ellas aumentó el volumen de contratación registrada. Cabe señalar, sin embargo, que para evaluar con mayor precisión los niveles de concentración de la actividad notarial deberíamos realizar un examen detenido de las regencias y otras situaciones particulares que, con una frecuencia indeterminada, podían implicar el mantenimiento nominal de una notaría monitorizada desde otra localidad, en la cual residía el notario y desde la cual es posible que realizara, si no la totalidad, una parte de los actos. Sabemos que en torno al 35% de los asientos mencionan alguna regencia y su influencia, por lo tanto, pudo ser notable.

**Figura 2: Censos consignativos registrados, por notaría**




Podemos ejecutar ejercicios analíticos similares tomando en consideración tipologías documentales específicas, en vez de la totalidad de los documentos registrados. En la figura 2 se ha cartografiado la actividad de las notarías en el que, en aquel momento, era el principal instrumento de crédito: los *censales* o censos consignativos. Si, como sostiene Postel-Vinay (1998), los notarios eran agentes básicos en la movilización del crédito privado, tiene mucho sentido tomarlos como referencia para analizar el funcionamiento del mercado del crédito; aunque cabe precisar que, en Catalunya,

buena parte del flujo crediticio procedía de las instituciones religiosas, especialmente el que se cedía mediante censos consignativos, y la red de penetración en el territorio de estas instituciones probablemente hacía menos indispensable el papel mediador de los notarios.

La cartografía de los censos consignativos arroja un resultado bastante distinto de lo observado para el conjunto de la actividad notarial (figura 2). El alcance de los contratos de crédito realizados fuera del distrito hipotecario se reduce y, lo que es más significativo, Barcelona desaparece como plaza de contratación. El flujo de crédito parece que se articulaba en un mercado más local, más cerrado. Por otra parte, aparece una jerarquía territorial bastante más marcada. Ciertamente, el volumen de contratación de censos consignativos es sólo una parte de la contratación total y, por lo tanto, hay un cambio de escala que no hemos amortiguado. Sin embargo, Girona aparece, en ambos periodos, como un lugar destacado. En 1768-1774 concentraba el 26% de las operaciones; y en 1775-1780, en un distrito más reducido, su peso aumentó hasta el 55%. La explicación se encuentra en las características de la oferta crediticia. Si bien en el mercado de los censos consignativos participaban un sinfín de instituciones religiosas, desde beneficios y causas pías, hasta colegios canonicales y monasterios, la mayor oferta procedía de las grandes instituciones que se acumulaban en las ciudades y, en este caso, en Girona. Sólo en el primer periodo aparece otro núcleo que compite en centralidad con la ciudad de Girona. Se trata de la villa de Castelló d'Empúries, antigua capital del condado homónimo y núcleo muy activo en el siglo XVIII, donde también existía una notable concentración de instituciones religiosas.

**Más allá de las notarías: Desafíos en la desambiguación de otros topónimos**

A pesar del papel que pudieron jugar las notarías en la articulación del mercado financiero, su cartografía no agota el examen de los flujos de crédito ya que sólo nos permite posicionar las notarías como espacios de contratación y mediación, pero nada nos dice sobre cada una de las partes contratantes, aunque en los párrafos anteriores hayamos propuesto algunas deducciones al respecto. La misma consideración puede aplicarse a cualquier otra tipología documental. Nuestro objetivo, pues, no puede limitarse a identificar las notarías, sino que debe extenderse a todos los topónimos presentes en cada asiento.

Ello plantea retos más complejos. Un topónimo puede aparecer en diversos momentos del documento y con funciones también diversas. Algunos indican el lugar de residencia de los actores, o su naturaleza; otros ubican cada finca objeto de la transacción; en los márgenes aparecen notas de traslado del asiento a los libros de otro oficio de hipotecas; etc. Muchos topónimos son nombres de lugares, municipios o parroquias, pero aquellos

que se utilizan para ubicar una finca en su paraje o que delimitan sus lindes son microtopónimos que, en la escala de análisis en que estamos trabajando, pueden tener poca utilidad. La función de cada topónimo puede ser determinada a partir de su posición respecto a palabras clave que se repiten sistemáticamente. Por ejemplo, cuando encontramos la secuencia "natural de dicha villa de $llocAbs[topónimo]" sabemos con certeza que su función es indicar la naturaleza del individuo citado pocas palabras antes. *Natural* o *naturales* son términos clave, como también lo son *residente*, *habitante*, *de presente en*, *oy en*, etc.

La proximidad a palabras clave permite descifrar parte de la función realizada por muchos topónimos, pero no su totalidad. Podremos conocer la naturaleza de un individuo, pero no si este es, por ejemplo, el comprador o el vendedor, el censalista o el censatario, lo cual es determinante para una exploración a fondo de las relaciones tanto sociales como territoriales. Se trata de informaciones que se refieren a conjuntos de ítems más que a un ítem concreto. En este caso, el dato concierne tanto al antropónimo como al oficio o a los lugares de naturaleza y residencia, y sólo puede determinarse examinando el documento en su conjunto. La gestión automática de este proceso exige que se realice por tipologías documentales. Mientras que, considerados conjuntamente, los asientos del registro siguen una multiplicidad de estructuras discursivas, si observamos cada tipología singularizadamente constatamos que la estructura interna es muy repetitiva. Los distintos elementos que la conforman mantienen un orden recurrente y, lo que es más interesante, se utilizan palabras clave para separarlos. En un censal, por ejemplo, se repiten los términos *creó* o *crearon* para distinguir al censatario del censalista. En este caso no importa tanto la proximidad como el orden (antes de/después de) para determinar la función de cada conjunto de ítems.

Finalmente, otra dificultad que se plantea en el proceso de desambiguación de los topónimos es el hecho que, frecuentemente, aparecen en cadenas jerárquicas donde coexisten con otros topónimos. Por ejemplo, en la siguiente venta el comprador aparece como domiciliado en la villa de Hostalric y en el obispado de Girona

> *... vendió al $tractament[magnifico] $antroponim[francisco rovira y de robles] Donzel en la* ***villa de $llocAbs[Hostalric] Obispado de $llocAbs[Girona]*** *domiciliado*

O en el siguiente, se menciona un vecindario y la parroquia donde de halla:

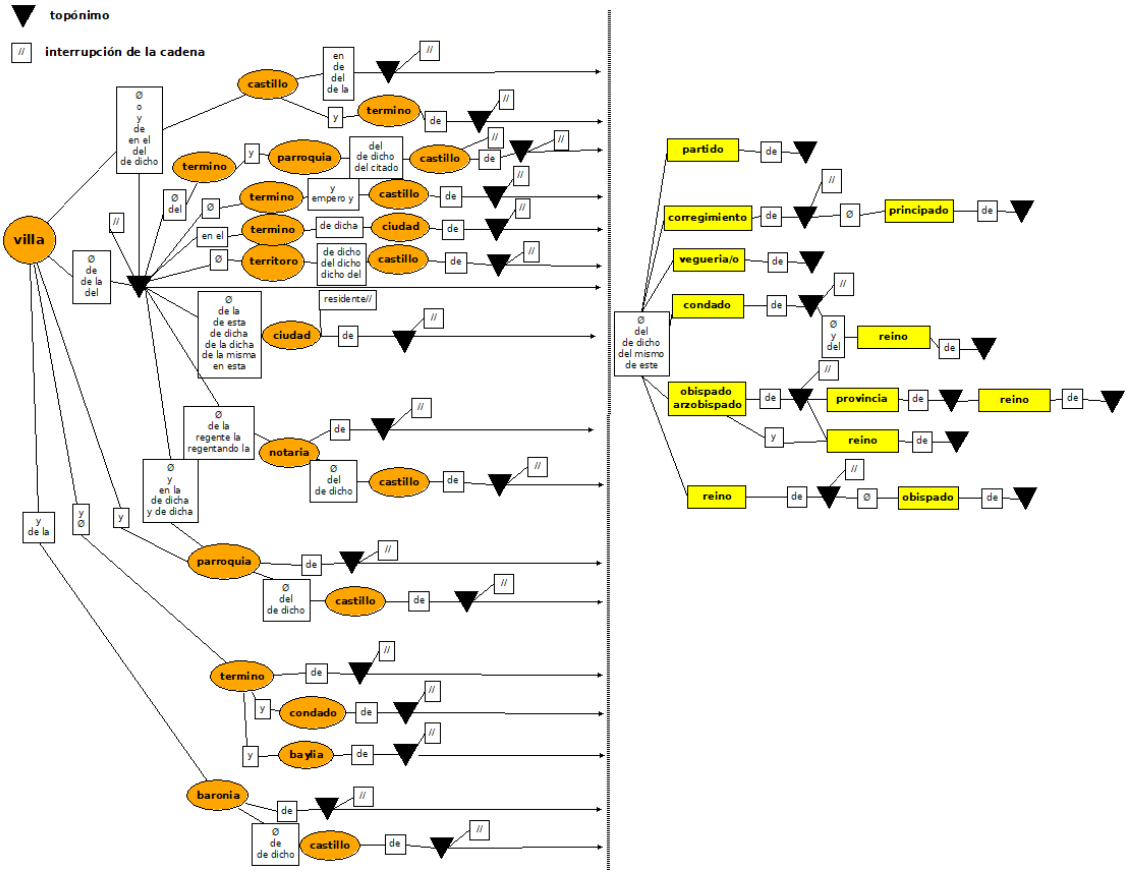
> *...establecieron á $antroponim[julian vila] $ofici[carretero] del* ***vezindado de las Barracas Parroquia de dicho Lugar de $llocAbs[Celrà]***

Se trata de una información valiosa porque, en muchas ocasiones, permite asignar unas coordenadas más precisas y, en otras, clarifica la localización cuando haya dudas o ambigüedad –como sucede con un topónimo del tipo *las Barracas*. Sin embargo, las cadenas jerárquicas de topónimos pueden llegar a ser bastante complejas, dando lugar

a múltiples combinaciones. Entonces, su simplificación operativa –la reducción mecánica a sus coordenadas– se torna más compleja.

En la figura 3 puede visualizarse el abanico de posibilidades que hemos hallado en todas las cadenas que se iniciaban con el término *vezindado*. Los términos clave (vezindado, castillo, parroquia, villa,...) que identifican cada tipo de unidad territorial aparecen en naranja cuando se trata de entidades locales o en amarillo cuando se trata de unidades administrativas superiores, ya sean civiles (partido, veguerío, corregimiento), feudales (condado) o eclesiásticas (obispado). Los topónimos concretos se han sustituido por un triángulo que los simboliza, y en un recuadro aparecen los términos de enlace que pueden encontrarse en cada paso. El resultado final, como puede observarse, es un número considerable de combinaciones posible, que aumenta geométricamente si consideramos que la cadena también puede iniciarse con otros términos (villa y lugar), en vez de vecindario.

**Figura 3: Cadena jerárquica de topónimos iniciada con el término *vezindado***



**Cartografía social de los flujos de crédito**

Si bien estamos trabajando para refinar el proceso de extracción y desambiguación de topónimos y hacerlo capaz de abordar con el rigor requerido la enorme variabilidad que presentan las cadenas toponímicas, hemos explorado, de manera provisional, estrategias alternativas cuyos resultados, aunque renuncian inicialmente a alcanzar la máxima precisión, pueden dar una idea bastante aproximada de las posibilidades que ofrece la extracción de los topónimos que indican la residencia de los contratantes.

Así, se ha desarrollado un algoritmo relativamente sencillo, programado en el lenguaje VBA (Visual Basic para Aplicaciones) que se halla integrado en la hoja de cálculo Excel, con el que se han analizado los contratos de creación de censos consignativos. En este tipo de contratos, como ya se ha dicho, determinadas palabras clave permiten localizar en el texto, con una razonable precisión, los dos fragmentos consecutivos en los que pueden encontrarse los datos relativos a censalistas y censatarios. Cada uno de los fragmentos incluye al menos una cadena toponímica (más de una, cuando el contrato se establece entre prestamistas o prestatarios múltiples, en cuyo caso se trabaja sobre la que aparezca en primer lugar) que, tras ser aislada, pasa a ser analizada por el algoritmo. Se identifican así las categorías toponímicas que la componen, incluidas las referencias a topónimos mencionados en partes previas del texto, y se selecciona la unidad administrativa más próxima al municipio, lo que permite asignar el código correspondiente.

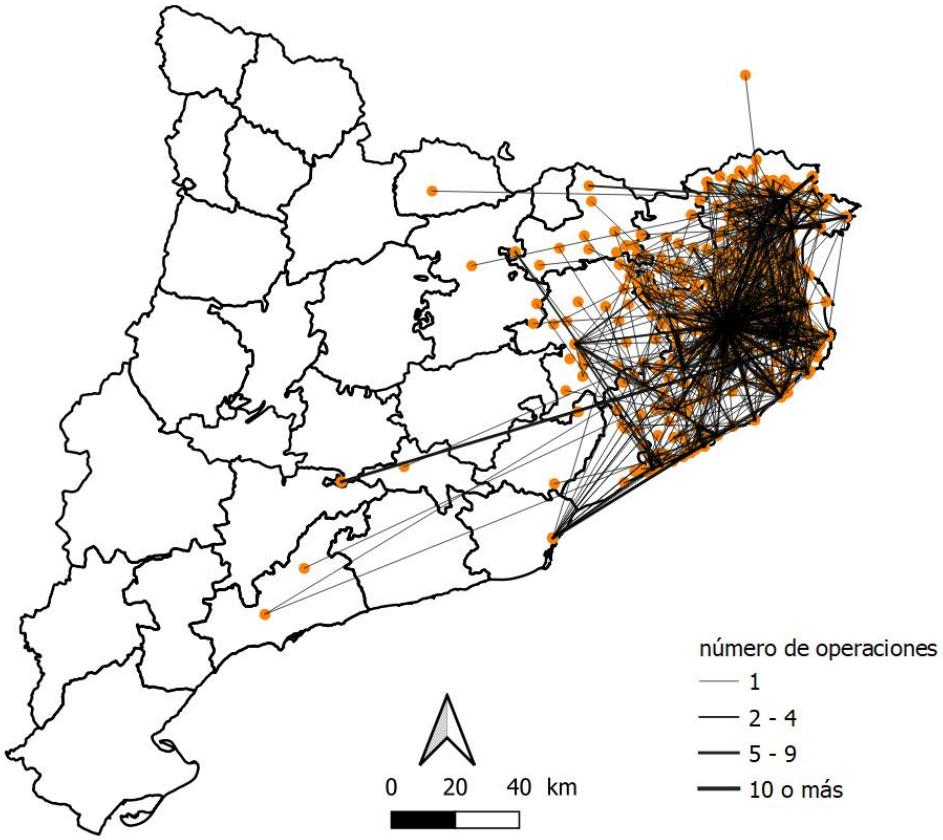
Debe advertirse que la utilización de estos datos residenciales, en el estadio de desarrollo actual del proyecto, requiere una tarea de supervisión y revisión manual bastante intensa. Por ello nos limitaremos al periodo 1768-1773, constituido por un total de 3.580 *censales*, porque es el bloque documental que ya se ha revisado.[2] Más adelante, para cerrar la comunicación, volveremos a la cuestión de la gestión de errores, su coste y sus retos.

Es sabido que el mercado del crédito en el siglo XVIII tenía un fuerte componente interpersonal, aunque en el caso catalán buena parte de él fluyera desde las instituciones eclesiásticas. No sólo su contratación se hallaba poco centralizada, en lo que se refiere a los lugares notariales. Tampoco lo estaban su oferta y, aunque esto ya fuera de esperar, su demanda. El análisis espacial de estos flujos mediante herramientas de GIS permite observarlos con detalle y profundizar en sus características. Para ello es importante conjugar simultáneamente los datos geográficos (toponímicos) que

[2] Aunque la mayor parte de las escrituras registradas en el Oficio de Hipotecas datan de pocos meses antes, también se inscribieron algunas hechas con bastante anterioridad. En 1771 el 12% de asientos correspondía a escrituras realizadas antes de 1768. Entre ellos podemos encontrar, incluso, escrituras del siglo XIII o XIV. Cabe indicar que todas las escrituras anteriores a la creación del registro han sido depuradas y eliminadas del análisis de censos consignativos que hemos realizado.

identifican tanto a quien ofrece como a quien toma a crédito, como sus identidades sociales, cuya extracción sigue unas pautas similares a lo expuesto hasta el momento.

**Figura 4. Flujos del crédito censalista, según residencia de las partes contratantes (1768-1773)**

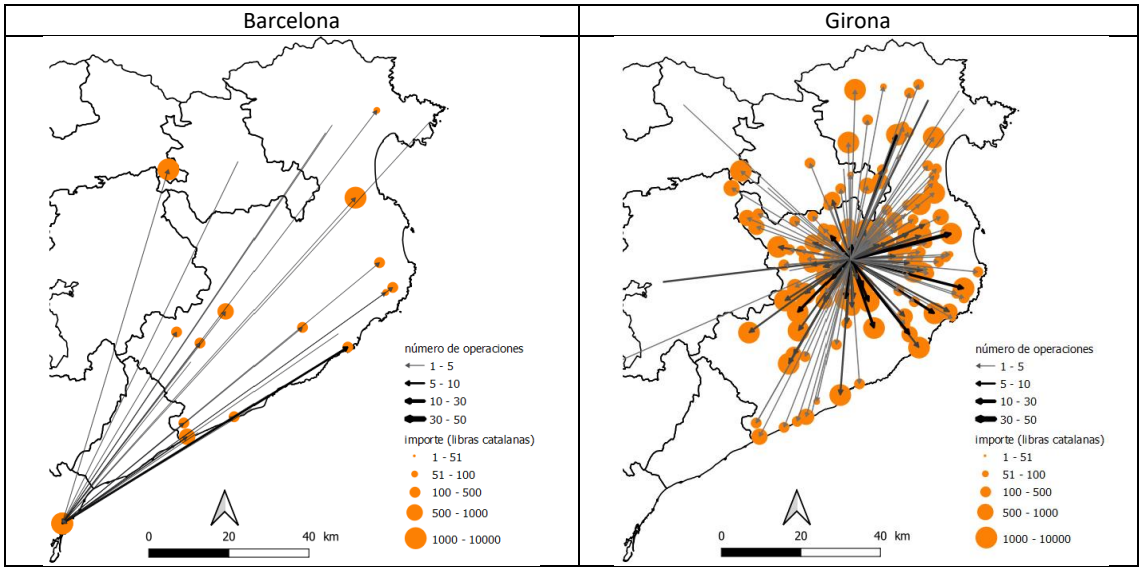



El primer paso conduce a la generación de un mapa de flujos general, que relacione todas las localidades de residencia de los ofertantes (censalistas) y de los demandantes (censatarios). El resultado, como se observa en la figura 4, es un amasijo de interrelaciones de tal densidad que hace difícil sacar conclusiones que vayan más allá de la constatación de que la mayor parte de las transferencias se realizaban dentro de los límites del propio distrito hipotecario. Algunos de los flujos singulares que aparecen en el mapa, y que atraen la atención por su carácter excepcional y porque marcan los límites de este mercado financiero local, se explican por razones y relaciones personales (soldados censalistas originarios del lugar de residencia del censatario, pastores pirenaicos que trashumaban hasta el litoral ampurdanés, apoderados que gestionaban patrimonios de propietarios foráneos, ...). Sin embargo, lo más relevante, lo que marca la pauta es que el radio en el que operaba el mercado local del crédito estaba básicamente circunscrito al propio distrito y las localidades próximas a él. La distancia media entre prestamista y acreedor era sólo de 10 kilómetros, y en un 35% de las veces, ambos residían en la misma localidad. Sólo el 1,5% de las operaciones indican una

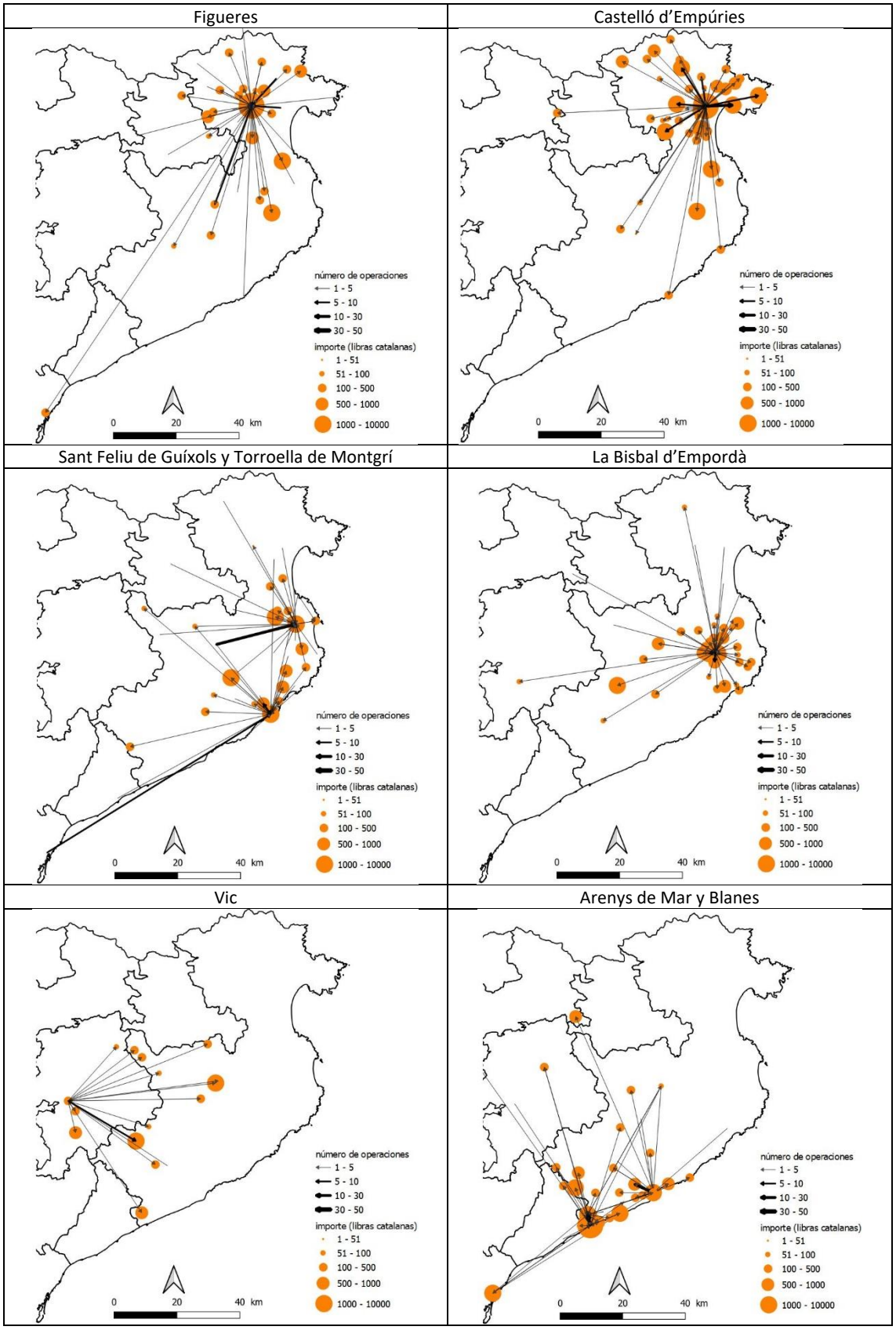
distancia superior a los 50 km. El mercado del crédito era un ámbito de confianza y de conocimiento cercano, lo cual atenuaba los riesgos inherentes a la actividad crediticia.

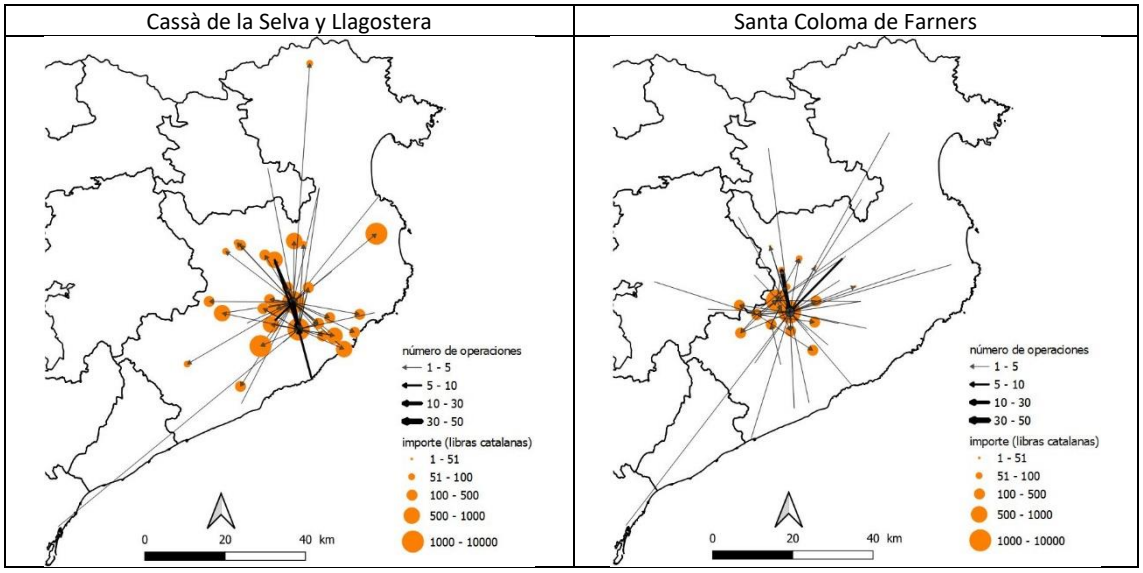
También puede deducirse del mapa de flujos general que, dentro del distrito gerundense, existían dos núcleos de mayor concentración, correspondientes a las ciudades de Girona y Figueres, aunque la densidad de las tramas no permite dilucidar, a primera vista, su peso y su sentido en términos globales. Para avanzar y para añadir otras capas de información sin deteriorar la legibilidad de la figura, es necesario descomponer los elementos de la madeja crediticia. A continuación, se reproducen algunos mapas correspondientes a los flujos generados en torno a villas y ciudades concretas. En ellas se superponen el capital tomado a censal por los residentes en cada lugar y procedente de la ciudad/villa seleccionada (mediante un círculo anaranjado escalable)[3], el sentido de las operaciones (mediante una flecha hacia el lugar de residencia de los censatarios) y la cantidad de operaciones (escalando el grosor de la misma flecha). Se ha utilizado el complemento mmqgis de QGis, que también permite obtener las distancias lineales entre cada uno de los nodos en que existe un flujo.

**Figura 5. Cartografía de los flujos de crédito, según lugar de residencia de censalistas y censatarios (1768-1773)**




[3] El círculo anaranjado sobre la ciudad o villa seleccionada en cada mapa agrega tanto el flujo de capitales procedentes de otros lugares prestado a vecinos de dicha ciudad o villa como el flujo interno de vecinos que prestaron a otros vecinos del mismo lugar.


Figueres
Castelló d'Empúries
Sant Feliu de Guíxols y Torroella de Montgrí
La Bisbal d'Empordà
Vic
Arenys de Mar y Blanes
número de operaciones
1 - 5
5 - 10
10 - 30
30 - 50
importe (libras catalanas)
1 - 51
51 - 100
100 - 500
500 - 1000
1000 - 10000
0
20
40 km




La secuencia de mapas permite destacar el papel de la ciudad de Girona como centro de oferta de crédito, aunque también pone de relieve la relevancia de otros centros secundarios que pasan desapercibidos en el mapa general. Si se observa con atención el sentido de los flujos, se puede constatar que, mientras en algunos dominan las indicaciones de oferta y salida de capital, otros se caracterizan por absorber capital, especialmente de los puntos más alejados. El caso extremo lo constituye la villa de Santa Coloma de Farners, cuyo mapa indica que su oferta de crédito se ceñía a localidades muy cercanas, al tiempo que atraía capital de lugares bastante más lejanos.

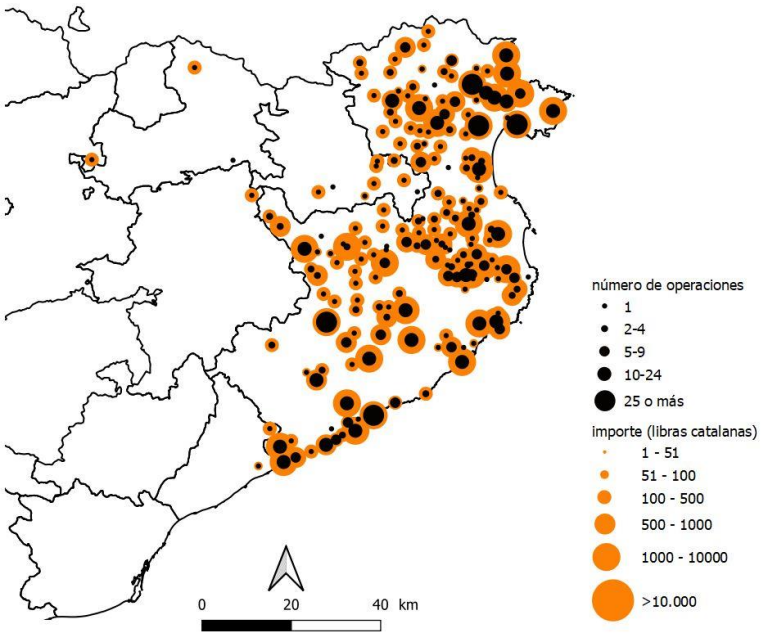
No es casual que, en la tabla 1, Santa Coloma de Farners aparezca en último lugar con la distancia media entre los censalistas que residían en ella y los censatarios que accedían a su crédito en sólo 2,7 km, en buena medida como consecuencia de que las tres cuartas partes de los censatarios residían en la milla villa que sus censalistas. En general los que hemos identificado como centros secundarios en la oferta de crédito se caracterizan por una elevada proporción de transacciones entre vecinos de la misma localidad, entre la mitad y las tres cuartas partes. Sólo un par de ellas, Cassà de la Selva y la Bisbal d'Empordà, se sitúan en una tercera parte. De conjunto, vuelve a destacar claramente la ciudad de Girona, con sólo un 13% de transacciones entre vecinos. También es la única localidad del distrito hipotecario (ni Vic, ni Barcelona forman parte de él) con un valor de distancia promedio (16 km) que supera claramente la media global.

**Tabla 1. Amplitud del mercado del crédito, según la distancia entre censalista y censatario**

| *lugar de residencia del censalista* | *núm. contratos* | *distancia promedio (km)* | *residencia en el mismo lugar* |
|---|---|---|---|
| Barcelona | 24 | 82,50 | 0% |
| Girona | 405 | 15,97 | 13% |
| Vic | 29 | 15,03 | 0% |
| Figueres | 66 | 10,81 | 50% |
| Cassà de la Selva | 93 | 8,63 | 31% |
| Castelló d'Empúries | 212 | 7,40 | 47% |
| la Bisbal d'Empordà | 106 | 6,20 | 34% |
| Blanes | 62 | 4,80 | 71% |
| Arenys de Mar | 71 | 4,52 | 63% |
| Sant Feliu de Guíxols | 78 | 4,23 | 69% |
| Torroella de Montgrí | 36 | 3,74 | 61% |
| Llagostera | 53 | 3,54 | 68% |
| Santa Coloma de Farners | 48 | 2,71 | 77% |

Uno de los objetivos básicos del proyecto de investigación, y el que nos animó en su momento a aventurarnos en la tarea de transcribir literalmente una gran colección documental, fue la posibilidad de visualizar de manera relativamente rica y compleja la actuación de aquellos grupos sociales que han dejado menor rastro documental y son, en consecuencia, más esquivos a la investigación histórica. Nos interesan particularmente aquellos que, en el siglo XVIII, se autoidentificaban como trabajadores; y nos interesa observar cómo, en unas condiciones de crecimiento demográfico y hambre de tierras, evolucionaron sus condiciones de vida y sus posibilidades de acceso a la tierra, y también al crédito. Su presencia en los protocolos notariales y, por elevación, en los registros de hipotecas era conocida, pero hasta el momento se ha aprovechado relativamente poco desde la perspectiva apuntada por la dificultad de manejo de las fuentes. A medio plazo, nuestro objetivo es disponer de una muestra amplia de trayectorias individuales de trabajadores, reconstruidas a partir de las actividades que se registraron en los libros del Oficio de Hipotecas. Hasta ahora hemos podido cuantificar su importante presencia en el mercado del crédito, básicamente como demandante y, con frecuencia, en operaciones vinculadas a la adquisición de derechos de propiedad sobre la tierra (establecimientos enfitéuticos, compraventas,...) (Congost, Garcia-Orallo y Saguer, 2021). Para finalizar esta exposición sobre la explotación cartográfica de la base generada con la transcripción de los libros de hipotecas, veamos qué nos puede indicar la cartografía de los trabajadores beneficiarios de un censal, según su lugar de residencia.

**Figura 6: Localización de los trabajadores que accedieron a crédito entre 1768 y 1773**




Cabe indicar que, en el periodo examinado, los trabajadores fueron los receptores del crédito en un 29,6% de todos los censales que se crearon.[4] Este porcentaje está algo por debajo de su peso demográfico en la estructura social,[5] pero no demasiado; por contra el volumen global de capital prestado, aunque los préstamos no fueran de poca monta para un trabajador del setecientos, sí que tienen un peso más bajo, del orden del 14%. En cualquier caso, los trabajadores –y esto es lo importante– tenía un nivel considerable de acceso al tipo de crédito hipotecario que constituían los censos consignativos. Tiene interés cartografiarlo para analizar si este acceso estaba generalizado o manifestaba pautas diferenciales en territorios distintos, aunque relativamente cercanos. Cuando lo hacemos (figura 6), confirmamos que la distribución territorial del acceso al crédito no era homogénea y que pueden distinguirse varias zonas. Era muy intenso en el alto Ampurdán, especialmente en la zona vitícola más septentrional, y flojeaba en la comarca de la Selva, a medida que se acercaba al Montseny. El cruce de estos datos con la cartografía de los establecimientos enfitéuticos, el principal instrumento que permitió tanto acceder a derechos individuales de propiedad sobre la tierra como disponer de bienes hipotecables para pedir crédito, es probable que aporte resultados interesantes en la comprensión de las dinámicas de expansión y crecimiento agrario del siglo XVIII.

[4] Entre 1768 y 1773 se crearon 3.580 nuevos censales, por un vaor total de 792.439 libras. Los trabajadores fueron beneficiarios de dichos préstamos en 1.061 operaciones, cuyo importe ascendió a 113.674 libras, que se corresponden con el 29,6% de las operaciones y el 7% del capital.

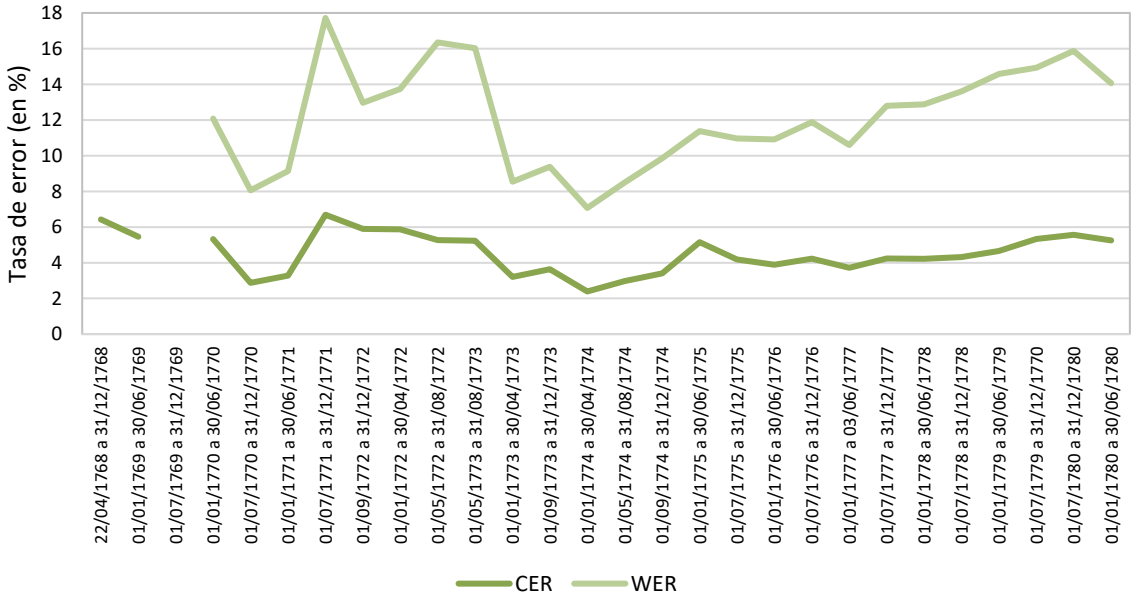
[5] En 1769 el peso de los trabajadores se situaba en un 33% de la estructura social. Ver Congost, Ros & Saguer, 2015.

**De vuelta a la transcripción: la gestión de los errores.**

Advertíamos anteriormente que la realización de un ejercicio como el que acabamos de desarrollar nos había exigido un proceso de depuración y revisión manual bastante intenso. Uno de los desafíos mayores de nuestro proyecto es alcanzar un equilibrio óptimo entre el coste marginal que supone la corrección de una multiplicidad de pequeños y muy variados errores, y la obtención de una muestra que, por su tamaño y calidad, pueda ser indiscutiblemente robusta.

Las métricas usadas para evaluar la calidad general de la transcripción arrojan unos valores que, en el estadio tecnológico actual, son muy aceptables. La tasa de error en caracteres (CER) es del orden del 4,5% de promedio. Ello indica que el 95,5% del texto, considerado carácter a carácter, se ha transcrito sin error. Si evaluamos el porcentaje de palabras que contienen algún error (WER), el valor crece hasta un promedio del 12,1%. Este aumento es esperable dado que el error en un sólo carácter afecta a toda la palabra y tiene, por ende, efectos multiplicadores.

**Figura 6: Tasas de error en la transcripción por volumen**




CER: tasa de error en caracteres; WER: tasa de error en palabras. Las fechas del eje de abcisas corresponden a cada uno de los libros del Oficio de Hipotecas de Girona.

Por otra parte, el resultado no es homogéneo a lo largo de la serie por distintos motivos. En primer lugar, por la distinta calidad de la escritura. Aunque precisamente una de las ventajas que ofrecen los libros del Oficio de Hipotecas de Girona es la uniformidad en el trazo, hay otros elementos que inciden sobre la calidad de les escritura, como las manchas y borrones, la transparencia del papel o la cantidad de tinta traspasada entre

una cara y su reverso. Cabe tener en cuenta, además, que la repentina mejoría de las métricas en el último volumen de 1773 coincide con un cambio en la estrategia empleada por los algoritmos encargados de la transcripción —en concreto, se pasó de utilizar un modelo estadístico, las cadenas ocultas de Markov, a un enfoque basado en redes neuronales—. Finalmente, la evolución de los resultados también obedece a una evolución del trazo respecto el modelo que fue utilizado para entrenamiento del sistema. Es probable que en algún momento debamos realizar tareas de reentrenamiento para evitar que el resultado siga empeorando como parece observarse en los últimos volúmenes.

Las métricas de transcripción se han obtenido testeando el resultado automático de una muestra de 50 imágenes seleccionadas al azar en cada uno de los volúmenes. Son de gran utilidad para controlar los resultados de la transcripción, pero tienen un carácter general. No indican nada sobre la calidad precisa de aquellos ítems de información que son más relevantes para nuestros objetivos, y este no es un detalle baladí ya que es precisamente en las palabras que se repiten con menor frecuencia, especialmente los topónimos y los antropónimos, donde pueden concentrarse más errores. El sistema transcriptor "aprende" a medida que se le entrena y se le corrige; lo cual comporta una tendencia a mejorar el resultado en los términos más frecuentes, y viceversa.

**Tabla 2. Correcciones referidas al topónimo identificador de la notaría**

| | parcial | total |
|---|---|---|
| Total documentos | | 67.426 |
| notaria no identificada por el transcriptor | | 10.925 |
| reconstruïdas con búsquedas selectivas y reemplazos masivos | 10.281 (15,2%) | |
| no reconstruïdas | 644 (1,0%) | |
| notaria identificada por el transcriptor | | 56.501 |
| identificación correcta y normalizada | 18.736 (27,8%) | |
| identificación corregida o normalizada | 36.964 (54,8%) | |
| identificación incorrecta | 801 (1,2%) | |

Volviendo al ejemplo inicial de las notarías, tal como se observa en la tabla 2, el número de notarías identificadas sin necesidad de introducir ninguna corrección ha sido sólo del 27,8%. La mayor parte de ítems ha requerido alguna enmienda en el proceso de normalización. Estas correcciones pueden realizarse con relativa simplicidad a partir de tablas de equivalencia de las variantes más frecuentes con la forma normalizada. Con ello alcanzamos un 82% de referencias correctas. Del resto, en un 16% el sistema no identificó al topónimo correspondiente a la notaría y se intentó reconstuirlo mediante búsquedas selectivas para, una vez seleccionados los registros pertinentes, reemplazarlos en lote. Ya se expuso anteriormente que el caso de las notarías es especial por su relativa simplicidad. Aun así, el manejo de miles de datos con algún tipo de error,

aunque sea fácilmente enmendable, exige un tiempo considerable cuando se pretende utilizar el corpus entero.

## Conclusión

La transcripción automática de textos manuscritos ha empezado a ser una realidad. Es probable que los avances que se realicen en los años venideros modifiquen sustancialmente la formas de trabajar con documentación manuscrita, como ya lo han hecho con la documentación impresa. También cambiarán las competencias de los investigadores, que deberán adquirir habilidades para aprender a manejar grandes cantidades de documentos transcritos. En el estadio actual, la explotación de textos transcritos automáticamente aún plantea muchos desafíos derivados tanto de los errores derivados de la transcripción como de las dificultades de extracción de los ítems de información y su conversión en datos procesables. La comunicación presentada ha pretendido ilustrar tanto estas dificultades como los procedimientos implementados para superarlas a partir de un ejemplo concreto referido a la toponimia. Este ejemplo también ha permitido plantear, a modo de ilustración, cómo podemos vincularlo a un sistema de análisis geográfico que contribuya a enriquecer el análisis que, en nuestro caso, hemos concretado en el mercado del crédito.

## Referencias

Bosch, V.; Congost, R.; Quirós, L.; Saguer, E.; y Vidal, E. (2018) “El reconocimiento automático de texto manuscrito aplicado a la Contaduría de Hipotecas de Girona”, *Transiciones en la agricultura y la sociedad rural. XV Congreso de la SEHA*, Santiago de Compostela.

Congost, R.; Garcia-Orallo, R. ; & y Saguer, E. (2021) “Seeing Credit and Property Rights from Below. The Experience of Catalan Smallholders in the Eighteenth Century”, en revisión.

Congost, R.; Ros, R.; y Saguer, E. (2015) “Beyond Life Cycle and Inheritance Strategies: The Rise of a Middling Social Group in an Ancien Régime Society (Catalonia, Eighteenth Century)”, *Journal of Social History*, 49 (3), 617–646.

Postel-Vinay, G. (1998) *La terre et l’argent. L’agriculture et le crèdit en France du XVIIIème au debut du XXème siècle,* París, Armand Colin.

Quirós, L.; Serrano, L.; Bosch, L.; Toselli, A.; Congost, R. Saguer, E. & Vidal, E. (2018). “Oficio de Hipotecas de Girona. A dataset of Spanish notarial deeds (18th Century) for Handwritten Text Recognition and Layout Analysis of historical documents”. Geneva: Zenodo (https://zenodo.org/record/1322666).

Quirós, L.; Bosch, L.; Serrano, L.; Toselli, A.; & Vidal, E. (2018), “From HMMs to RNNs: Computer-Assisted Transcription of a Handwritten Notarial Records Collection”, *ICFHR 2018 The 16th International Conference on Frontiers in Handwriting Recognition*, Niagara Falls (EUA).