ALMA MATER STUDIORUM

UNIVERSITÀ DI BOLOGNA · SEDE DI CESENA

FACOLTÀ DI SCIENZE MATEMATICHE, FISICHE E NATURALI

Corso di Laurea Magistrale in Scienze e Tecnologie Informatiche

Analisi e integrazione dati finalizzate alla lotta digitale all'evasione fiscale

Tesi di Laurea in Sistemi Informativi

Relatore:

Chiar.mo Prof.
Matteo Golfarelli

Presentata da:

Enrico Gallinucci

Correlatore:

Dott. Camillo Acerbi

III Sessione

2011/2012

Indice

Indice delle figure

Indice delle tabelle

Introduzione

Negli ultimi anni, il problema dell'evasione fiscale è stato fortemente discusso sul territorio nazionale. I dati del 2012 mostrano come l'Italia sia il Paese con il maggiore tasso di evasione in Europa, nonché tra i più alti dell'area OCSE. A risentire di questo problema non è solo lo Stato ma anche le amministrazioni comunali, alle quali spettano parte degli incassi ricavati dalla tassazione. Inoltre, la recente crisi economica e le conseguenti politiche di austerità adottate dal Governo hanno fatto sentire il loro peso sui bilanci dei Comuni, che si trovano ad avere sempre meno risorse da destinare alle loro attività. In questo difficile scenario economico la necessità di combattere efficacemente l'evasione fiscale è diventata più che mai importante. Grazie all'evoluzione tecnologica, questa problematica può essere oggi affrontata anche dal punto di vista informatico: incrociando le informazioni memorizzate negli archivi digitali di vari enti è infatti possibile individuare comportamenti o situazioni che testimoniano un'evasione fiscale. In questa direzione si è già mossa l'Agenzia delle Entrate con la recente messa in opera di Serpico ("Servizio per i contribuenti"), un applicativo in grado di incrociare una moltitudine di dati per scovare gli evasori. Anche a livello locale cominciano a nascere iniziative volte a combattere l'evasione fiscale. E' in questo contesto che si è concretizzato il progetto su cui si focalizza questa tesi, denominato DyNamiTE (Digital fightiNg Tax Evasion) e svolto dall'Università di Bologna in collaborazione con il Comune di Cesena.

L'obiettivo principale di questo progetto è quello di sviluppare uno strumento informatico che consenta di individuare situazioni di sospetta evasione fiscale nel territorio di Cesena, con particolare attenzione alla casistica degli affitti in nero. Il primo passo necessario sarà quello di analizzare le banche dati e integrarle in un

unico schema. Successivamente bisognerà studiare i dati a disposizione per definire e implementare le tecniche di ricerca degli evasori; allo staff dell'Ufficio Tributi del Comune verrà inoltre richiesto un supporto nella verifica dei risultati e nella validazione delle tecniche implementate. Al fine di rendere disponibile lo strumento realizzato al personale del Comune sarà infine necessario lo sviluppo di un'interfaccia web, che permetta di consultare agevolmente i dati integrati e di applicare le tecniche di ricerca degli evasori.

La realizzazione di un nucleo informativo che racchiude e integra le informazioni di banche dati diverse non è necessariamente confinata alla lotta all'evasione fiscale, ma può essere sfruttata in diversi contesti. Un esempio riguarda la possibilità di eseguire analisi di tipo previsionale (*what-if*) in funzione di determinati scenari. Un ultimo obiettivo del progetto è quindi quello di concretizzare queste analisi e dimostrare il potenziale informativo della banca dati integrata.

L'esposizione del lavoro svolto in questa tesi viene suddivisa in quattro capitoli. Nel primo viene approfondita la problematica da affrontare e vengono descritte la struttura del progetto e le banche dati a disposizione. Nel secondo capitolo viene fornita una panoramica delle principali tecnologie utilizzate per la realizzazione del progetto. Nel terzo si entra invece nel vivo del progetto, con la documentazione di tutte le attività svolte per lo sviluppo del back-end del sistema, costituito dalla banca dati riconciliata. Infine, il quarto e ultimo capitolo si focalizza sulla parte di front-end: essa comprende l'implementazione dei pattern di evasione, lo sviluppo dell'interfaccia web e la realizzazione di un modello previsionale focalizzato sull'equità fiscale.

1 Il Sistema Informativo del Comune di Cesena

La ricerca degli evasori fiscali parte dall'analisi complessiva del sistema informativo del Comune di Cesena; in primo luogo è necessario approfondire la problematica da affrontare, per capire qual è la situazione di partenza e da dove nasce la lotta digitale all'evasione. Successivamente si entra nel merito del progetto descrivendo tutte le attività previste, dalla costruzione dello schema riconciliato agli utilizzi che di esso si vuole fare. Un ultimo paragrafo viene infine dedicato alle singole banche dati, con lo scopo di fornire una panoramica generale dei dati impiegati nel corso del progetto.

1.1 La problematica

Il sistema informativo del Comune di Cesena dispone di una serie di banche dati, gestite internamente o importate da enti terzi, che spaziano su diverse aree di interesse. Sebbene queste banche dati siano fisicamente memorizzate nello stesso sistema informativo, i relativi schemi risultano sconnessi tra loro, indipendenti; in questo modo, i dati contenuti in schemi diversi - oltre a non essere consultabili contemporaneamente - non hanno alcun legame con i dati degli altri schemi. Teoricamente parlando, è però naturale che le varie banche dati presentino delle aree di sovrapposizione, dei concetti comuni in cui compaiono pressoché le stesse informazioni. Ad esempio, la banca dati dell'anagrafe e quella del catasto trovano un punto di intersezione nei soggetti memorizzati: l'elenco dei residenti e l'elenco dei titolari catastali, seppur presentino differenze dal punto di vista sintattico, condividono in gran parte gli stessi soggetti (i titolari degli immobili risiedono mol-

to spesso nello stesso comune in cui si trovano i rispettivi immobili). Se si uniscono i due elenchi, le informazioni anagrafiche e catastali di ogni persona diventano anch'esse unite tra loro e si stabilisce un vero e proprio legame tra le due banche dati.

La possibilità di integrare i dati di banche dati diverse fornisce due importanti vantaggi. Il primo riguarda la velocità delle operazioni di consultazione. Quando gli schemi sono integrati tra loro, la navigazione delle informazioni di banche dati diverse diventa semplice e immediata; ad esempio, per conoscere l'indirizzo di residenza di un titolare catastale serve al massimo qualche secondo. Il secondo vantaggio riguarda invece il notevole potere informativo acquisito. All'interno delle singole banche dati, le informazioni sono limitate a un ruolo principalmente operativo, di basso livello (registrazione e semplice consultazione di eventi o di proprietà). Quando invece si prendono tutte insieme, queste stesse informazioni acquisiscono un ruolo più strategico: esse diventano infatti parte di complessi profili, permettendo di disegnare storie o scenari da interpretare. Ad esempio, il titolare catastale può diventare un single di mezza età che, dopo la morte del padre, ha affittato la casa ereditata a una coppia di ragazzi stranieri, e così via. L'estrazione e l'interpretazione di queste storie può così essere sfruttata per individuare situazioni strane o apparentemente incongrue; un esempio volutamente banale consiste in una titolarità catastale valida per una persona che risulta deceduta. Le incongruità nelle situazioni rilevate possono essere dovute a errori manuali o a segnalazioni tardive - come sarebbe probabile nell'esempio precedente. In altri casi, tuttavia, le incongruenze potrebbero essere reali e volontarie, ad indicare una possibile circostanza fraudolenta.

E' in questo contesto che prende forma la **lotta digitale all'evasione fiscale**: essa ha infatti l'obiettivo di sfruttare l'unione dei dati per individuare le suddette circostanze fraudolente, in particolare quelle in cui emergono sufficienti dati per supporre che un determinato soggetto abbia perseguito un'evasione fiscale.

1.2 La struttura del progetto

In questo paragrafo si vuole delineare l'impostazione generale del progetto, dalla costruzione dello schema riconciliato agli utilizzi che di esso si vuole fare. Lo svolgimento complessivo del progetto viene infine riassunto brevemente nell'ultimo paragrafo.

1.2.1 L'integrazione delle banche dati

Come anticipato nel paragrafo precedente, la maggior parte delle banche dati di cui dispone il Comune di Cesena gode di una propria indipendenza rispetto alle altre. Si tratta cioè di banche dati non integrate, separate tra loro e memorizzate in schemi diversi. Ad esempio, i dati gestiti e controllati dall'anagrafe comunale fanno riferimento a uno schema che non ha nulla a che vedere con quello del catasto, nonostante esistano delle naturali sovrapposizioni nelle informazioni gestite.

La mancanza di integrazione delle banche dati comporta due problemi fondamentali. Il primo consiste nell'impossibilità di interrogare contemporaneamente banche dati diverse; in questo caso, la soluzione è piuttosto semplice: l'importazione di tutte le informazioni in un unico schema basterebbe per rendere i dati completamente disponibili. Tuttavia, il secondo problema risulta quello più complicato: assunto il fatto che nelle diverse banche dati siano presenti dei concetti comuni tra loro (come i soggetti, gli indirizzi, ecc.), l'indipendenza dei relativi schemi fa sì che tra tali concetti non ci sia alcuna relazione. Si riprenda l'esempio delle persone fisiche, per cui i soggetti che possiedono almeno una proprietà catastale nel Comune di Cesena siano spesso residenti nello stesso Comune. Da questa relazione si intuisce come l'elenco dei titolari catastali sia composto perlopiù da soggetti già presenti in anagrafe, ma la mancanza di integrazione degli schemi fa sì che le stesse informazioni risultino scollegate e ripetute sia nella banca dati dell'anagrafe che in quella del catasto.

Se quindi si riunissero tutte le informazioni dalle varie fonti in un unico schema, mancherebbero comunque le relazioni tra i concetti comuni delle rispettive banche dati. Di conseguenza, non solo le informazioni risulterebbero duplicate, ma le stesse ripetizioni potrebbero risultare incongruenti tra loro: in assenza vin-

coli, elementi teoricamente identici tra loro (una stessa persona, o uno stesso indirizzo) possono essere espressi in maniera più o meno diversa nelle varie banche dati. Un indirizzo può essere scritto secondo formalismi diversi, o un errore manuale può causare la memorizzazione di dati leggermente diversi per un soggetto, come una data di nascita diversa o - ancora peggio - un cognome diverso. La presenza di questo di tipo di differenze fa sì che l'individuazione delle associazioni tra le informazioni sia complicata: se si considerassero i soli *match perfetti*, una fetta rilevante delle associazioni reali andrebbe persa.

Per ovviare a questo tipo di problema si rende necessario l'utilizzo di tecniche di **join approssimato**, le quali permettono di riconoscere i collegamenti anche in assenza di corrispondenze perfette. Date due tabelle che esprimono un concetto comune, ogni record dell'una viene confrontata con i record dell'altra, calcolando una misura di distanza su uno o più campi; terminati i confronti, a ogni record della prima tabella viene associato il record dell'altra tabella con cui la distanza risulta minima. I risultati di tutti le associazioni vengono quindi memorizzati in apposite strutture, dette *tabelle di match*. In questo modo si riescono a creare le connessioni tra i concetti comuni di schemi diversi, concretizzando l'integrazione delle diverse banche dati e dando vita a un unico schema riconciliato su cui poter effettuare le indagini richieste.

1.2.2 L'obiettivo principale: la ricerca degli evasori

Terminata l'integrazione delle banche dati a disposizione, lo schema riconciliato può essere sfruttato per attuare la lotta digitale all'evasione fiscale. L'attività principale consiste quindi nel definire i cosiddetti *pattern di evasione*, ossia i percorsi di navigazione dello schema riconciliato che, attraverso la selezione di dati con specifiche caratteristiche, restituiscano un elenco di situazioni sospette - ordinate, ove possibile, con un grado di sospetto decrescente.

In termini generali, l'evasione fiscale può spaziare su diversi settori, per cui è necessario stabilire innanzitutto su quale tipo di frode puntare. In seguito ad un'analisi delle banche dati a disposizione e a una valutazione delle attività di lotta all'evasione già portate avanti dall'Agenzia delle Entrate, si è deciso di concentrar-

si principalmente sulla lotta agli **affitti in nero** - senza comunque escludere la possibilità di estendere la ricerca ad altri tipi di evasione.

Per la definizione di un pattern di evasione esistono fondamentalmente due tipi di approcci: il primo consiste nel definire i criteri tipici dello scenario legale e individuare i casi in cui questi criteri non sono rispettati. Un esempio semplificato potrebbe prevedere la rilevazione di un sospetto nei casi in cui un residente abiti in un appartamento non di sua proprietà (né di proprietà di un parente) e non sia registrato un contratto d'affitto tra il residente stesso e il proprietario dell'appartamento. Si tratta di casi teoricamente semplici da individuare, ma che richiedono un ottimo livello di qualità dei dati, in cui tutte le situazioni legali siano correttamente individuabili e in cui non ci siano problemi di dati mancanti. Un secondo approccio consiste invece nel mettersi nei panni dell'evasore, capire quali siano gli escamotage tipicamente adottati per nascondere un affitto in nero e definire i criteri con cui tali situazioni possono essere rilevate nei dati a disposizione. Un esempio riguarda le false separazioni: due coniugi dichiarano residenze separate (due prime case), quando invece convivono in una di esse e sfruttano la seconda casa per gli affitti. Si tratta di casi più complessi, in cui la difficoltà principale consiste nel definire in maniera corretta ed esaustiva il profilo dell'evasore, senza rischiare di includere situazioni con caratteristiche simili ma legalmente valide.

In entrambi i casi, è bene sottolineare come questo tipo di indagini non siano in grado di rilevare le irregolarità con certezza assoluta, ma permettano solamente di sollevare dei sospetti: possono essere svariati i motivi a giustificazione di situazioni anomale, primo tra tutti l'errore umano nell'inserimento dei dati. Al fine di massimizzare le probabilità di correttezza al sollevamento di un sospetto, è importante che tali ricerche siano il più precise possibile. Tuttavia, ogni situazione di potenziale evasione rilevata deve essere manualmente controllata e validata, sia per verificare la correttezza del pattern utilizzato, sia per escludere eventuali falsi positivi. Solo al termine di questi controlli, i nominativi selezionati possono essere effettivamente segnalati alle autorità competenti per gli accertamenti - ed eventualmente le sanzioni - del caso.

1.2.3 Gli utilizzi alternativi dello schema riconciliato

La ricerca degli evasori fiscali costituisce l'obiettivo principale di questo progetto, ma lo schema riconciliato dispone di potenzialità che vanno oltre questa singola attività; con i dovuti accorgimenti è possibile sfruttare in maniera alternativa lo schema riconciliato, in supporto a progetti già esistenti o programmati dal Comune di Cesena, oppure individuando funzionalità completamente nuove. Per cominciare, il Comune di Cesena ha in programma lo sviluppo del "**Fascicolo del cittadino**", ovvero uno strumento che riassuma tutte le informazioni sui residenti e che metta a disposizione tali informazioni ai dipendenti comunali. In tale contesto, la banca dati integrata di questo progetto si inserirebbe molto facilmente, costituendo essa stessa una buona parte del lavoro richiesto. Un esempio di nuova funzionalità lo si può invece trovare nella possibile applicazione di tecniche di **data mining**. La vastità di informazioni messa insieme nello schema riconciliato può essere sfruttata per ricercare caratteristiche comuni tra i cittadini; si potrebbe quindi cercare di raggruppare la popolazione in gruppi (detti *cluster*) in cui si possano evidenziare caratteristiche o comportamenti condivisi.

Per dimostrare le potenzialità della banca dati integrata, una parte del progetto è stata dedicata allo sviluppo di una funzionalità alternativa alla lotta digitale all'evasione fiscale. La scelta è ricaduta nell'utilizzo della banca dati in senso predittivo piuttosto che conoscitivo: ciò che si vuole fare è implementare un meccanismo in grado di simulare degli scenari futuri a partire dai dati attuali, per estrarre informazioni che possano essere di supporto al Comune nella scelta di politiche future. In particolare, le simulazioni si incentrano sul calcolo del carico fiscale dei residenti in termini di IMU e di addizionale IRPEF: modificando i parametri su cui il Comune ha facoltà di scelta, si vuole determinare il gettito totale previsto della tassazione e l'impatto che queste modifiche hanno sulle famiglie, opportunamente suddivise in categorie. L'obiettivo ultimo è quello di effettuare valutazioni sull'**equità fiscale** tra le tipologie di famiglie e di capire come le modifiche ai parametri delle tasse possa far variare - ed eventualmente migliorare - il grado di equità. Di questo meccanismo previsionale verrà soltanto realizzato un prototipo dimostrativo, una *proof-of-concept* (POC): non si vuole implementare una vera e

propria funzionalità, quanto semplicemente dare dimostrazione concreta delle potenzialità della banca dati integrata.

1.2.4 Le fasi del progetto

La realizzazione del progetto DyNamITE si può essenzialmente suddividere in quattro fasi. La prima fase riguarda l'integrazione delle banche dati in un unico schema riconciliato. Si comincia quindi con la scelta dei dati da prelevare dalle sorgenti e con la progettazione dell'architettura dello schema, che deve poi essere popolato con una serie di operazioni ETL (Extraction, Transformation and Loading). Successivamente vengono realizzati i join approssimati per agganciare le varie parti dello schema: si individuano i concetti comuni che richiedono il join approssimato, si determinano le regole di match e si implementano le procedure.

Terminata l'integrazione dello schema, la seconda fase prevede l'implementazione dei pattern di evasione. Viene quindi studiato lo schema per capire quali percorsi possono essere perseguiti, dopodiché si eseguono dei test per valutare la bontà degli stessi. Per determinare la correttezza dei risultati, è richiesto al personale dell'Ufficio Tributi del Comune di effettuare un riscontro manuale, sulla base del quale validare il pattern o modificarne i parametri di ricerca.

Una terza fase prevede invece la realizzazione di un'interfaccia grafica da mettere a disposizione dei dipendenti comunali. Le funzionalità dell'interfaccia consistono principalmente nella libera navigazione dello schema integrato, nell'esecuzione dei pattern validati nella fase precedente e nella possibilità di confermare manualmente i match approssimati determinati nella fase di integrazione.

Nella quarta e ultima fase viene infine realizzato il POC sull'equità fiscale: viene quindi progettato il meccanismo previsionale e si implementano le procedure per il calcolo dei carichi e fiscali, dopodiché vengono simulati alcuni scenari su cui eseguire una serie di analisi.

1.3 Le banche dati

In questo paragrafo vengono fornite informazioni generali sulle banche dati utilizzate e integrate in questo progetto, come la qualità dei dati e gli utilizzi previsti. Le banche dati vengono raggruppate in base alle rispettive fonti di provenienza.

1.3.1 Comune di Cesena

La principale fonte è ovviamente costituita dallo stesso Comune di Cesena, il quale detiene le banche dati primarie su cui basare la costruzione dello schema riconciliato. La banca dati più importante viene sicuramente individuata nell'**anagrafe**, in cui sono contenute tutte le informazioni anagrafiche sui residenti a partire dal 1999 in poi. Importanti informazioni legate ai dati anagrafici sono anche gli indirizzi di residenza e i rapporti di parentela tra le persone. Qualitativamente parlando, i dati dell'Anagrafe godono di un buon livello; sono rari gli errori manuali e si possono trovare alcune incongruenze tra le date di immigrazione/emigrazione, ma si tratta di una banca dati complessivamente affidabile. Per questo motivo, i dati anagrafici possono essere considerati come il fulcro dello schema riconciliato: le operazioni di integrazione delle altre banche dati considereranno sempre i dati anagrafici come il riferimento su cui basare le operazioni di match.

L'altra banca dati fondamentale è costituita dalla **toponomastica**, in cui è definito il cosiddetto *stradario*: ogni via del Comune di Cesena viene definita con il suo toponimo ufficiale e associata ad un codice unico e ad un DUG (Denominativo Urbanistico Geografico). Inoltre, la toponomastica specifica l'elenco di tutte le abitazioni in termini di civico, civico-bis, interno e interno-bis. Nonostante il suo contributo informativo sia limitato, questa banca dati ha una grossa importanza nel definire ufficialmente e univocamente i toponimi delle strade, i quali vengono spesso scritti secondo i più variegati formalismi. Il ruolo della toponomastica è quindi quello di rappresentare il punto di riferimento su cui basare i match sugli indirizzi stradali. L'unica banca dati che non ha bisogno di questa operazione è quella dell'anagrafe, in cui gli indirizzi di residenza sono già conformi alla codifica toponomastica.

1.3.2 Agenzia delle Entrate

Nella costruzione del database integrato, l'Agenzia delle Entrate ricopre un ruolo fondamentale in quanto fornitrice di un'ampia gamma di banche dati. La prima ad essere considerata in questo progetto è quella del **catasto**, di cui il Comune di Cesena dispone già da diverso tempo e che viene scaricata mensilmente dall'Agenzia del Territorio (oggi incorporata nell'Agenzia delle Entrate). In questa banca dati sono memorizzate tutte informazioni storicizzate sugli immobili posizionati all'interno dei confini comunali. Oltre ai dettagli sui singoli immobili (come la categoria di appartenenza e la superficie), essa specifica tutte le informazioni sulle titolarità catastali (come le quote e la durata del possesso) e sulle persone detentrici di titolarità (sia fisiche che giuridiche). Inoltre, ad ogni immobile sono associate le coordinate catastali e gli indirizzi stradali. Il ruolo di questa banca dati è di grande importanza per la lotta digitale all'evasione fiscale: i collegamenti con l'anagrafe per determinare proprietari e residenti di un dato immobile sono spesso fondamentali. Purtroppo, i dati in essa contenuti soffrono di uno scarso livello qualitativo su diversi punti. Innanzitutto, le informazioni sui titolari catastali contengono spesso errori e uno stesso titolare può comparire più volte con alcuni dati leggermente diversi. Inoltre, le quote di possesso delle titolarità catastali non sono spesso specificate e la somma delle stesse non corrisponde sempre al 100%. Le note più dolenti riguardano tuttavia gli indirizzi: innanzitutto, i toponimi delle strade sono scritti senza usare alcun formalismo, per cui i riscontri con la toponomastica risultano difficili. In secondo luogo, non è prevista la memorizzazione dei civici interni degli immobili; questa grave mancanza fa sì che conoscere i residenti in un determinato appartamento risulta molto difficile, talvolta addirittura impossibile.

La seconda banca dati ricevuta è quella relativa alle **utenze dei consumi**, ossia le bollette di acqua, luce e gas, lei cui informazioni vengono inviate all'Agenzia delle Entrate dai vari gestori delle utenze (Enel, Hera, Edison, ecc.) e conglobate per tipologia di utenza (elettricità, acqua e gas). Richiesti dal Comune di Cesena specificamente per questo progetto, i suddetti dati risultano già inseriti nel sistema informativo e parzialmente integrati, almeno per quanto concerne la riconciliazione tra gli intestatari delle utenze e i residenti anagrafici. Sebbene le utenze ricevute

siano relative solamente all'anno 2010, tali informazioni solo importanti per la possibilità di ricavare una stima del numero di occupanti effettivi rispetto ai residenti (o affittuari) dichiarati. Purtroppo, i dati ricevuti non risultano completi: le informazioni sugli intestatari delle utenze, i tipi di utenza ed i consumi fatturati sono qualitativamente buoni, ma - come per le titolarità catastali - l'indirizzo sui cui è attiva l'utenza risulta privo del civico interno. L'associazione di un'utenza ad un appartamento risulta pertanto possibile nei soli casi in cui l'intestatario dell'utenza risieda nello stesso stabile.

Un contributo importante arriva inoltre dai dati sulle **dichiarazioni dei redditi**, anch'essi già inseriti nel sistema informativo del Comune e relativi agli anni 2006, 2008, 2009 e 2010. Informazioni utili includono il reddito totale, il reddito da fabbricati e il reddito imponibile. Tuttavia, si tratta di dichiarazioni solamente riassuntive: i redditi sono espressi solo come totali, senza il livello di dettaglio che il sito dell'Agenzia delle Entrate offre per le ricerche puntuali da parte degli operatori comunali. Tali dettagli includono, ad esempio, le specifiche del reddito da fabbricati per ogni immobile: tipologia dell'immobile, rendita catastale, uso dell'immobile ed eventuale reddito da locazione. Queste informazioni sarebbero particolarmente utili, nonché in grado di sopperire alle mancanze del registro dei contratti d'affitto. Per questo motivo, il Comune ha sottoposto all'Agenzia delle Entrate la richiesta di ricevere i dati completi (come accade già per i maggiori Comuni d'Italia); tuttavia, ad oggi tale richiesta non è stata ancora soddisfatta. Le dichiarazioni riassuntive restano comunque valide ed utilizzabili; da esse è possibile sapere quale sia la ricchezza complessiva di una famiglia, oppure confrontare il reddito totale da fabbricati di una persona con la rendita catastale delle sue proprietà per effettuare dei controlli di congruenza.

L'ultima banca dati ricevuta dall'Agenzia delle Entrate è quella relativa ai **contratti d'affitto**, contenente l'elenco dei contratti depositati nell'ufficio di Cesena (relativamente agli anno 2009 e 2010), insieme alle informazioni sui rispettivi locatari, affittuari ed immobili coinvolti. Diversamente dalle altre, questa banca dati non era ancora stata inserita nel sistema informativo del Comune, perciò le informazioni sono disponibili solamente nei tracciati testuali scaricati dall'Agenzia delle Entrate. L'apporto informativo dei contratti d'affitto nella lotta all'evasione

fiscale è sicuramente notevole, essendo quest'ultima principalmente focalizzata nel combattere gli affitti in nero. Tuttavia, si rilevano una serie di problemi che limitano fortemente l'utilizzo di questi dati. Innanzitutto, ai tre quarti dei contratti non è associato alcun immobile, per cui l'espressività di tali contratti si riduce di molto. Inoltre, i contratti possono essere consegnati in un qualunque Comune, che può non essere quello in cui si trova l'immobile; ne risulta che, dei contratti "sopravvissuti" al passo precedente, la metà di questi riguarda immobili che non si trovano a Cesena. Ipotizzando che le stesse percentuali si presentino negli altri comuni, si può simmetricamente stimare che la metà dei contratti relativi ad immobili di Cesena vengano depositati in comuni diversi. La somma di questi due fattori fa sì che i dati effettivamente utili siano solo una minima parte; di conseguenza, i contratti d'affitto non possono essere usati per individuare attivamente delle evasioni (come veniva inizialmente ipotizzato), quanto solo per escludere eventuali falsi positivi.

1.3.4 ISTAT

In seguito al **21° censimento nazionale**, avvenuto il 9 ottobre 2011, l'istituto nazionale di statistica ha inviato ai Comuni d'Italia i tracciati dei dati raccolti; in particolare, le persone censite sono state suddivise in 3 categorie: "residenti censiti", "residenti non censiti" e "censiti non residenti". Per la lotta all'evasione fiscale, la categoria dei "censiti non residenti" è stata valutata interessante: essa contiene infatti le persone (con i rispettivi domicili) che hanno dichiarato di abitare a Cesena ma che non risultano residenti nel comune stesso - ossia potenziali affittuari su cui effettuare delle verifiche. Tuttavia, così come nel catasto e nelle utenze, gli indirizzi di residenza risultano senza interni e non conformi ai toponimi ufficiali; gli stessi problemi e le stesse operazioni di bonifica risultano necessarie anche in questa situazione.

2 Le tecnologie

La realizzazione del prototipo DyNamiTE ha richiesto l'adozione di diverse tecnologie nelle varie fasi del progetto:

- Uno strumento ETL in grado di popolare lo schema integrato importando i dati dalle varie sorgenti.
- Un RDBMS per ospitare la banca dati integrata.
- Un'interfaccia che consenta agli operatori comunali di usufruire delle funzionalità sviluppate.

I paragrafi seguenti descrivono in maniera dettagliata le caratteristiche e gli utilizzi di ciascuna tecnologia adottata.

2.1 Pentaho Data Integration

Pentaho Data Integration (PDI, conosciuto anche come Kettle) è un'applicazione client desktop dedicata precisamente all'integrazione dei dati, ossia all'esecuzione di operazioni ETL (Extraction, Transformation and Loading) [PEN13]. Realizzata dalla Pentaho Corporation, PDI fa parte di una suite di prodotti di Business Intelligence open-source chiamata Pentaho Business Analytics, la quale fornisce anche servizi OLAP, strumenti di reportistica e di data mining. I moduli della suite sono sviluppati in Java e sono disponibili per i tre principali sistemi operativi; inoltre, seppur integrati tra loro, i moduli della suite possono essere installati separatamente, a seconda delle proprie necessità.

La versione utilizzata di PDI è la 4.2.0 (Community Edition), ovvero l'ultima disponibile al momento dell'installazione. L'impiego di questo software è stato fon-

damentale per la creazione della banca dati riconciliata; grazie a PDI è infatti possibile eseguire le seguenti operazioni.

- Estrazione di dati da qualunque fonte (o sorgente): è possibile importare tabelle da qualunque RDBMS (MySQL, Oracle, MS Access, ecc.) o leggere dati da fonti di altri tipi (come fogli Excel o tracciati testuali). Data l'eterogeneità delle banche dati che devono essere integrate, la possibilità di importare dati da diverse tipologie di sorgenti è sicuramente importante.
- Trasformazione dei dati importati: una volta caricati, è possibile eseguire sui dati una vasta serie di operazioni; esse possono spaziare dai classici operatori SQL (join, raggruppamenti, eliminazione di duplicati) all'esecuzione di codice procedurale scritto dall'utente. Queste operazioni permettono di selezionare i dati che devono essere portati sulla banca dati riconciliata e operare su di essi una serie di bonifiche.
- Caricamento dei dati: il passo conclusivo consiste nell'inserire (o aggiornare) i dati in una banca dati di destinazione. Come per la lettura dei dati, anche la scrittura può avvenire sulle stesse tipologie di contenitori (tabelle, fogli Excel o file di testo). Nel nostro caso, le scritture avvengono esclusivamente sulla banca dati riconciliata in MySQL.

L'interfaccia grafica di PDI (denominata *Spoon*) consente di organizzare le operazioni sopraelencate in procedure ETL, chiamate *Trasformazioni* ed espresse attraverso dei grafici. Ogni tipologia di operazione è rappresentata da un'icona; trascinandola nell'area del grafico, essa può essere poi opportunamente configurata. Il flusso dei dati da un'operazione all'altra viene invece rappresentato da frecce direzionate, che collegano le varie icone posizionate.

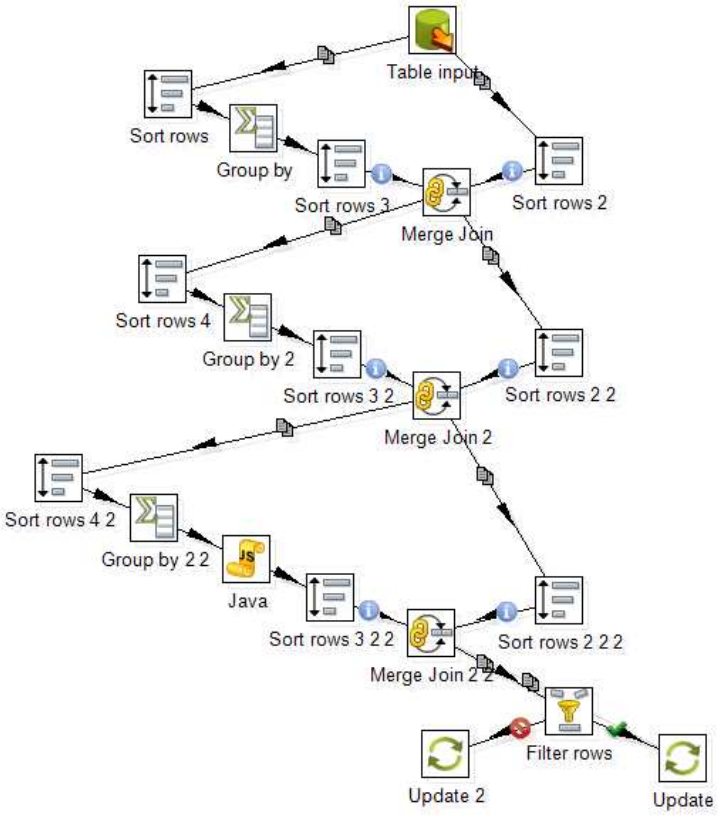

Figura 1 - Un esempio di Trasformazione PDI implementata

Le Trasformazioni costruite possono essere salvate ed eseguite in qualunque momento. Inoltre, due o più Trasformazioni possono essere collegate tra loro ed eseguite in sequenza; tale costrutto viene definito come *Job*. Sebbene i Job non siano mai stati utilizzati nel progetto, il loro impiego potrebbe essere valutato nell'ottica di automatizzazione delle procedure di costruzione e aggiornamento della banca dati integrata.

2.2 MySQL

MySQL è un RDBMS (Relational DataBase Management System) open source, sviluppato da Oracle Corporation e distribuito dalla stessa sotto licenza GPL (GNU General Public Licence). MySQL spicca per essere oggi l'RDBMS open source più diffuso al mondo [ORA13a], nonché adottato da famosi prodotti di calibro internazionale quali Wikipedia, Google, Facebook e tanti altri [WIK13a]. Tra i motivi del successo di questo software va citata la sua ampia portabilità: essendo scritto in C e C++, MySQL è utilizzabile su tutti i sistemi operativi che dispongano di un compilatore C++ [ORA13b]. Inoltre, MySQL è un componente fondamentale degli am-

piamente diffusi pacchetti software di tipo AMP (Apache, MySQL e PHP/Perl/Pyton), utilizzati per lo sviluppo di applicazioni web e disponibili per tutti i principali sistemi operativi (LAMP per Linux, WAMP per Windows, ecc.). Nel progetto DyNamiTE, MySQL - installato alla versione 5.5 Community Edition - è stato utilizzato per ospitare, manipolare ed interrogare la banca dati riconciliata. Ciascuna di queste operazioni ha richiesto l'utilizzo della GUI (Graphic User Interface) ufficiale, denominata MySQL Workbench e solitamente integrata nell'installazione. Tale interfaccia si suddivide in 3 pannelli principali.

- **SQL Development**. Questo è il pannello principale di MySQL: selezionando una delle connessioni disponibili (per collegarsi all'istanza server locale o ad un'istanza remota) si ha accesso ai database memorizzati sull'istanza selezionata. SQL Development è fondamentalmente dedicato all'esecuzione di codice SQL, sia attraverso la scrittura manuale del codice nelle schermate di editing, sia attraverso l'utilizzo di apposite maschere (le quali generano il codice corrispondente in automatico). L'utilizzo di questo pannello è stato intensivo fin dalle prime fasi del progetto. Durante l'integrazione delle banche dati è stato utilizzato per la creazione dei vincoli referenziali nello schema riconciliato, per la bonifica di alcuni dati e per l'esecuzione dei join approssimati. Nelle fasi successive è stato invece principalmente impiegato per l'esecuzione delle complesse interrogazioni elaborate (dai pattern di evasione al calcolo del carico fiscale) e per la definizione di statistiche sui dati a disposizione.
- **Data Modeling**. Il pannello di Data Modeling è un'interfaccia grafica che mette a disposizione dell'utente le funzionalità di *forward-engineering* e *reverse-engineering*. La prima consiste nel definire i modelli delle tabelle richieste e demandare ad una procedura automatica la creazione dello schema del database a partire dai modelli precedentemente definiti. La seconda - al contrario - prevede di creare il modello logico di un database esistente. La funzionalità di back-engineering è stata l'unica utilizzata, con lo scopo di creare automaticamente lo schema ER del database.

- **Server Administration**. MySQL memorizza le banche dati su appositi *database server*, detti anche *istanze* server, in modo che questi possano essere acceduti anche da remoto. Il pannello Server Administration permette di gestire le suddette istanze: le operazioni più comuni sono la gestione degli utenti e dei rispettivi permessi, la modifica dei parametri di configurazione del server e il controllo del file di log. In questo progetto, la presenza contemporanea di due sviluppatori (vedi paragrafo 3.1) ha richiesto di creare un'istanza server nella macchina di uno sviluppatore e di predisporre sull'altra un'istanza client che si collegasse al server. Al termine della fase di co-sviluppo, la macchina ospitante il client è stata dismessa ed è rimasto operativo soltanto il server.

2.3 PHP e il framework interno

PHP è un linguaggio di scripting server-side e open-source, utilizzato principalmente nello sviluppo web per la creazione di pagine dinamiche e di applicazioni web. Creato da Rasmus Lerdorf come semplice strumento di supporto alla sua homepage personale, PHP è oggi portato avanti dal PHP Group ed è il software open-source più utilizzato dalle aziende, nonché impiegato sul 75% dei siti web in cui il linguaggio server-side adottato viene specificato [W3T13a]. Insieme a MySQL, PHP costituisce un'importante componente nei pacchetti di tipo AMP (vedi paragrafo precedente) ed è anch'esso utilizzato da prodotti di fama internazionale, quali Wikipedia e Facebook. Sebbene esistano diversi editor per la scrittura di codice PHP (quali Aptana, PHPEclipse o Netbeans, tutti gratuiti), si è deciso di utilizzare il classico Notepad++, soprattutto per la sua semplicità di utilizzo. PHP è stato utilizzato nel progetto per lo sviluppo dell'applicazione web da mettere a disposizione dei dipendenti comunali. In particolare, l'impiego di PHP riguarda le necessità di interagire con il database, di presentare a video i risultati delle interrogazioni e di mantenere in sessione alcune informazioni utili per l'utente.

La scelta di PHP rispetto a soluzioni alternative (come, ad esempio, JSP) è stata dettata dalla possibilità di utilizzare un framework sviluppato internamente ai Sistemi Informativi del Comune e già utilizzato per la realizzazione di diversi applicativi ad uso interno. Tale framework utilizza PHP nella versione 5.2 e mette a dispo-

sizione una serie di funzionalità che semplificano e velocizzano lo sviluppo di un'applicazione web.

- Controllo centralizzato ed automatico degli accessi: il framework implementa un pannello amministrativo che consente di gestire i permessi di accesso per ogni utente alle varie applicazioni del Sistema Informativo. Grazie a questa gestione centralizzata, nella homepage di ogni utente vengono direttamente forniti i link alle applicazioni a cui egli ha accesso. Inoltre, il meccanismo di autenticazione degli utenti risulta uniformato e viene automaticamente incluso nelle applicazioni all'atto della loro creazione.
- Standardizzazione degli accessi al database: oltre ad utilizzare lo standard ODBC per garantire la connettività i più comuni DBMS, i risultati delle interrogazioni vengono automaticamente formattati in un array multidimensionale o in una tabella HTML personalizzabile.
- Layout predefinito: a tutti i componenti principali delle pagine HTML viene impostato un layout di base, in modo da uniformare il *look-and-feel* delle applicazioni create.

Per utilizzare il framework in un'applicazione non sono richieste complicate configurazioni, ma risulta sufficiente includere alcuni file PHP nelle pagine dell'applicativo.

2.4 Javascript

Javascript è un linguaggio di scripting client-side, anch'esso utilizzato nello sviluppo web per la gestione dell'interazione dell'utente con la pagina e la modifica dinamica dei contenuti della pagina. Sviluppato dalla Netscape Communications Corporation e dalla Mozilla Foundation, Javascript ha conosciuto un largo successo con la diffusione delle tecniche AJAX (Asynchronous JavaScript and XML), le quali prevedono la comunicazione asincrona di dati tra client e server [MDN13]. L'utilizzo di Javascript si è oggi esteso anche al mondo non-web: diverse applicazioni (quali Adobe Reader, Adobe Photoshop o la suite OpenOffice) consentono

infatti l'inserimento di script personalizzati grazie all'inclusione di un interprete Javascript.

Come PHP, anche Javascript è stato utilizzato per lo sviluppo dell'applicazione web, anche se con un ruolo di minore importanza. Il suo impiego è stato infatti mirato alla gestione delle comunicazioni asincrone (a rendere più piacevole la navigazione dell'utente) e per alcune semplici animazioni nei componenti delle pagine. A supporto di quest'ultima necessità si è deciso di fare affidamento a JQuery, ovvero la libreria Javascript ad oggi maggiormente diffusa [W3T13b]. Open-source e già integrata nel framework interno, JQuery semplifica infatti l'esecuzione di codice Javascript, dal recupero di elementi del DOM all'applicazione di una vasta gamma di effetti grafici [TJF13].

3 Il progetto DyNamiTE: il back-end

La prima parte del progetto è stata dedicata allo sviluppo del "cuore" del sistema, ossia la banca data riconciliata, attraverso l'integrazione delle banche dati d'origine. La prima parte del capitolo riassume le attività che sono state svolte nei progetti precedenti; una descrizione approfondita del loro lavoro è consultabile nei documenti di Tesi degli studenti che vi hanno preso parte [PAV12] [SOR12]. La seconda parte del capitolo è dedicata ai match approssimati, un'attività precedentemente iniziata e conclusa insieme al mio contribuito. La terza e ultima parte, incentrata sull'estensione del database riconciliato, è stata invece interamente svolta in autonomia dal sottoscritto.

3.1 Sintesi dei progetti precedenti

I progetti precedenti si sono concentrati principalmente sulla selezione, integrazione e bonifica delle prime banche dati. Per la creazione del database riconciliato (denominato "evasione"), la scelta delle fonti iniziali da cui estrarre i dati è ricaduta su quelle gestite dal Comune (anagrafe e toponomastica) e alcune di quelle fornite dall'Agenzia delle Entrate (catasto e utenze dei consumi). Tutte e quattro sono reperibili dal Sistema Informativo del Comune, il quale utilizza una piattaforma Oracle per la memorizzazione dei dati.

Per ciascuna banca dati vengono forniti di seguito i dettagli sulla qualità dei dati, sulle informazioni effettivamente importate e sulle eventuali operazioni compiute su di essi.

3.1.1 Anagrafe

I dati anagrafici dei residenti del Comune di Cesena costituiscono il punto di partenza per la costruzione del database riconciliato, in quanto contengono informazioni qualitativamente buone su tutti i soggetti di interesse. In particolare, le informazioni importate sono le seguenti:

- Elenco dei residenti e dei relativi dati anagrafici. Si tratta di dati non storicizzati, perciò le informazioni specificate per un residente corrispondono a quelle attuali e non c'è traccia delle eventuali precedenti modifiche.
- Gli indirizzi di residenza dei residenti. Sebbene esista uno storico completo degli indirizzi di residenza (i quali sono maggiormente soggetti a variazioni rispetto ai dati anagrafici), i dati importati esprimono solamente la fotografia delle residenze al 31 dicembre 2010. Inoltre, gli indirizzi sono espressi secondo il formalismo della toponomastica, anche se manca il vincolo referenziale con gli indirizzi toponomastici.

Come detto più volte, i dati anagrafici dei residenti sono tra i migliori dal punto di vista qualitativo; per questo motivo, tali dati sono stati importati in maniera "grezza", ossia senza apportare alcuna modifica. L'unica bonifica effettuata ha avuto luogo sugli indirizzi di residenza (per implementare il vincolo referenziale con gli indirizzi toponomastici), ma è stata applicata solo dopo l'introduzione degli indirizzi di residenza storicizzati (vedi paragrafo 3.3.1).

3.1.2 Toponomastica

Il compito fondamentale dei dati toponomastici è quello di rappresentare il punto di riferimento per la definizione dei toponimi ufficiali delle strade, ossia delle rispettive nomenclature corrette e univoche. Le informazioni importate corrispondono alle seguenti:

- Elenco e decodifica dei Denominatori Urbanistici Generici (DUG): ad esempio "VLE" per "viale", "CSO" per "corso", ecc.
- Elenco di tutte le strade del Comune di Cesena, con rispettivi toponimi e DUG.

- Elenco di tutti i possibili indirizzi, con le specifiche numerazioni civiche esterne ed interne in ogni strada del Comune.

Qualitativamente ottimi, i dati toponomastici sono stati importati senza essere minimamente modificati.

3.1.3 Catasto

Il catasto contiene informazioni di vario tipo riguardanti gli immobili siti nel Comune di Cesena: dalle semplici caratteristiche degli immobili, alle quote delle titolarità di possesso e alle coordinate catastali. A differenza dell'anagrafe e della toponomastica, le informazioni sono generalmente più complesse, per cui la loro comprensione può non essere immediata. Per questo motivo, la spiegazione dei dati importati viene trattata con maggior approfondimento.

- Elenco di tutte le Unità Immobiliari Urbanistiche (UIU) del territorio di Cesena e delle rispettive caratteristiche. Si tratta degli elementi di riferimento su cui si fonda il database del catasto e comprendono tutti i tipi di immobile: residenziali e commerciali, magazzini, capannoni e fabbricati in costruzione. Dal momento che le caratteristiche delle UIU possono essere soggette a diversi cambiamenti (aumento del numero di vani, rivalutazione della rendita catastale, ecc.), esse sono necessariamente storicizzate. Va inoltre citato il fatto che i terreni sono memorizzati in una struttura a parte; esulando però dallo scope principale del progetto, si è deciso di non importarli e di mantenere le sole UIU.
- Elenco non storicizzato di tutti i titolari catastali. Così come per la distinzione UIU/terreni, anche la distinzione tra persone fisiche e giuridiche è materializzata nel database del catasto con l'utilizzo di due strutture separate. In maniera analoga, si è deciso di importare i soli dati delle persone fisiche, sulle quali si prevede di concentrare le analisi.
- Elenco storicizzato delle titolarità catastali, le quali associano i titolari (fisici e giuridici) alle proprietà (UIU e terreni) specificando la quota, la durata e la tipologia del possesso. Diversamente rispetto ai titolari e alle proprietà, le titolarità sono memorizzate indistintamente in un'unica struttura e, in fase di importazione, si è deciso di mantenerle tutte: ad

esempio, può essere interessante sapere che un immobile è posseduto per metà da una persona fisica e per metà da una giuridica, oppure che una determinata persona possieda anche un certo numero di terreni.

- Elenco degli indirizzi in cui si trovano le UIU; tali indirizzi non sono associati direttamente alle UIU ma alle rispettive versioni storiche. Il motivo di questa particolarità risiede nel fatto che un immobile può subire un cambio di indirizzo: ad esempio, ad un appartamento situato all'angolo in un incrocio può essere spostato l'ingresso da una strada all'altra.
- Elenco delle coordinate catastali delle UIU; sebbene anch'esse siano associate alle singole versioni degli immobili, in fase di importazione si è deciso di associarle direttamente alle UIU. Il motivo risiede nel fatto che i cambiamenti di coordinate catastali sono molto meno frequenti; pertanto, per ogni immobile si è deciso di impostare come unicamente valida la coordinata catastale più recente.

Dal punto di vista qualitativo, i dati catastali presentano una serie di problemi che devono essere necessariamente trattati. In primo luogo, gli indirizzi degli immobili sono scritti secondo i più variegati formalismi, rispettando solo in pochi casi i toponimi ufficialmente dichiarati in toponomastica. Parallelamente, i dati dei titolari presentano spesso degli errori di ortografia o di duplicazione di stesse persone. Questi problemi sono stato affrontati con l'applicazione delle tecniche di join approssimato (vedi paragrafo 3.3). In secondo luogo, l'assenza della numerazione civica interna negli indirizzi limita notevolmente l'integrazione della banca dati riconciliata, impedendo di associare con sicurezza un indirizzo di residenza ad un immobile. Per questo problema non esiste purtroppo una soluzione informatica o algoritmica: l'unica possibilità consiste in una bonifica manuale dei dati con l'aggiunta delle informazioni mancanti.

Un ultimo problema è quello della somma delle quote di possesso di un immobile (in determinati momenti o fasce temporali), la quale risulta spesso incongrua: può infatti capitare che essa sia diversa dal 100% (fenomeno causato il più delle volte dall'assenza delle quote di possesso nelle titolarità) o che non sia rispettata l'esatta corrispondenza tra le quote di nuda proprietà e quelle di usufrutto. In

questi casi, si è deciso di introdurre dei *flag* che indichino quando le due condizioni non sono rispettate, in modo da sapere se le quote di possesso di un determinato immobile siano o meno affidabili.

3.1.4 Utenze dei consumi

Provenienti dalle banche dati dei vari gestori sul territorio di Cesena, le utenze contengono informazioni sui consumi di elettricità, acqua e gas relativamente all'anno 2010. Memorizzate in tabelle separate a seconda del consumo, le informazioni contenute risultano pressoché identiche: intestatario dell'utenza, consumo fatturato, importo dovuto e indirizzo dell'abitazione. Al momento dell'importazione dei dati, l'integrazione tra gli intestatari delle utenze ed i residenti anagrafici era già stata implementata; rimane comunque necessaria l'integrazione tra gli indirizzi nelle utenze e gli indirizzi toponomastici, per la quale si rimanda al paragrafo 3.2.2. Tuttavia, similmente agli indirizzi catastali, anche gli indirizzi delle utenze risultano privi della numerazione civica interna, impedendo anche in questo caso la possibilità di associare un'utenza ad un indirizzo di residenza nel caso esistano più civici interni.

3.1.5 Schema dati

Al termine dei progetti precedenti, lo schema dati del database riconciliato era quello indicato in Figura 2. Nel modello logico sono riportate anche le tabelle di match (corrispondenti alle tabelle su sfondo bianco), di cui si parla nel paragrafo 3.2. Inoltre, per motivi di spazio, sono state omesse le tabelle che contengono solamente le decodifiche degli acronimi e delle sigle utilizzate in altre tabelle. Infine è necessario sottolineare che nello schema riconciliato è stata originariamente adottata la dicitura "enel" per contraddistinguere le utenze relative ai consumi di energia elettrica; sebbene i gestori delle utenze siano in verità diversi, tale dicitura è stata mantenuta negli sviluppi di questo progetto per mantenere l'uniformità dei formalismi utilizzati.

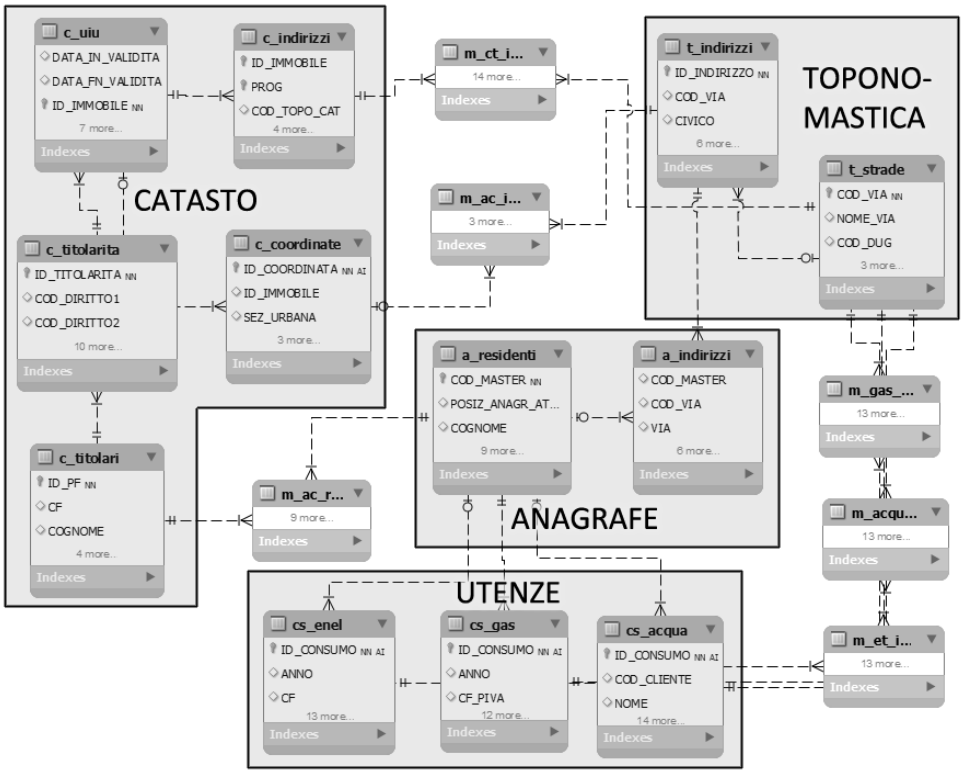

Figura 2 - Architettura iniziale del database riconciliato

3.2 Match approssimati

L'integrazione dei dati provenienti da banche dati diverse non si raggiunge con la semplice importazione delle informazioni in un unico database: un'operazione di fondamentale importanza consiste nell'applicazione delle tecniche di join approssimato, grazie alle quali risulta possibile eseguire i collegamenti tra i concetti comuni presenti. In questo progetto, sono due i concetti condivisi dalla maggior parte delle banche dati e su cui è richiesta l'applicazione di queste tecniche: le persone fisiche e gli indirizzi.

Il problema di base consiste nell'individuare le associazioni tra i record di due tabelle i cui elementi sono concettualmente identici; ad esempio, la tabella dei titolari catastali e quella dei residenti contengono entrambe persone fisiche, le quali risultano spesso ripetute sia nell'una che nell'altra tabella. L'individuazione delle associazioni è resa difficile dal fatto che, nelle diverse tabelle, una stessa persona può essere espressa in maniera diversa: che sia per un errore ortografico o per una comunicazione sbagliata, i valori dei campi nelle due tabelle potrebbero essere diversi; in questi casi, l'utilizzo dell'operatore di uguaglianza per il confronto dei

record non sarebbe in grado di effettuare il collegamento. Per risolvere questo problema è necessario fare affidamento a specifici algoritmi che permettano di riconoscere la similarità di due record attraverso l'utilizzo di misure di distanza (o di similarità).

Il meccanismo di join approssimato basa il confronto tra i record delle due tabelle su una serie di regole progressive, le quali specificano diversi livelli di tolleranza sulla distanza tra i due record. La progressività delle regole indica una decrescente rigidità: se i risultati del confronto rientrano nei parametri di una regola, l'associazione viene riconosciuta e memorizzata; in caso contrario si passa alla regola successiva, i cui parametri avranno una maggiore tolleranza. Terminate le regole, l'associazione viene naturalmente scartata. L'utilizzo di regole progressive permette inoltre di associare ai match individuati un diverso grado di certezza: minore è il numero della regola, maggiore è la sicurezza che il join individuato sia corretto. Tipicamente, la prima regola individua i match perfetti, per i quali è richiesta una completa uguaglianza su tutti i campi. Le regole successive individuano invece i match approssimati, per i quali vengono indicati i valori massimi accettati nelle distanze tra i campi.

Per misurare la differenza tra campi numerici o booleani, il calcolo della distanza è semplice: si considera una distanza 0 se i valori sono uguali, 1 se sono diversi. Nel caso di confronti tra campi alfanumerici, invece, si è deciso di adottare la distanza di Levenshtein, nota anche come distanza di edit [WIK13b]: date due stringhe a e b, la distanza è uguale a $\mathrm{lev}_{a,b}(|a|,|b|)$ dove:

$$\mathrm{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & , \min(i,j) = 0 \\ \min \begin{cases} \mathrm{lev}_{a,b}(i-1,j)+1 \\ \mathrm{lev}_{a,b}(i,j-1)+1 \\ \mathrm{lev}_{a,b}(i-1,j-1)+[a_i \neq b_j] \end{cases} & , \text{ else} \end{cases}$$

Figura 3 - Formula matematica della distanza di Levenshtein tra due stringhe

In altre parole, la distanza si calcola come il numero di operazioni elementari necessarie per trasformare la stringa a in b; per "operazione elementare" si intende l'aggiunta o la rimozione di un carattere, oppure la sostituzione di un carattere con un altro). Una volta calcolata, la distanza di Levenshtein viene poi normalizzata, ovvero viene rapportata alla lunghezza della prima stringa, in modo che la di-

stanza rientri in un valore compreso tra 0 (uguaglianza) e 1 (massima disuguaglianza).

Il calcolo della distanza si complica se si considerano stringhe composte da una sequenza di parole. In questi casi, il mero confronto tra le due stringhe potrebbe dare risultati fuorvianti: ad esempio, se la stringa *b* contiene tutte le parole di *a* in ordine inverso, la distanza di Levenshtein risulterebbe molto alta nonostante l'errore sia sul solo ordinamento delle parole. Il meccanismo di confronto viene quindi maggiormente elaborato per evitare il suddetto problema. Dalle due stringhe vengono innanzitutto estratte le singole parole, separate tra loro dal carattere " " (*spazio*); questa operazione viene definita *tokenizzazione*, dove ad ogni parola estratta corrisponde un *token*. Successivamente, ogni token della stringa *a* viene associato al token della stringa *b* con cui la distanza è minima. La distanza totale viene quindi determinata come la somma pesata delle distanze tra i singoli token. Infine, lo stesso procedimento viene ripetuto invertendo le due stringhe e si determina la media delle due distanze calcolate.

L'implementazione degli algoritmi fin qui descritti non può essere frutto di pure query SQL, bensì è richiesto l'utilizzo di un linguaggio quantomeno procedurale. Tra le varie opzioni disponibili si è deciso di sfruttare il linguaggio C#, il cui compilatore è integrato nel framework .NET, disponibile con Windows 7 e utilizzabile gratuitamente. Gli algoritmi di match approssimato sono stati pertanto implementati in appositi programmi C# e compilati da linea di comando, in modo da creare un file eseguibile lanciabile in qualunque momento.

Un'ultima osservazione riguarda la memorizzazione dei match: tutti gli agganci riconosciuti devono infatti essere salvati in apposite tabelle, in modo da materializzare e concretizzare l'integrazione dei due schemi di partenza. Seppur con qualche differenza in base alle singole implementazioni, le tabelle di match presentano tutte la stessa struttura.

- Chiave primaria, composta dall'identificativo del record proveniente dalla prima tabella e dall'identificativo del record proveniente dalla seconda tabella.
- Valori delle distanze tra i due record e/o tra i singoli campi.

- Numero progressiva della regola che ha determinato l'aggancio tra i due record; si considerano le regole ordinate per importanza decrescente, con la numero 1 ad indicare un match perfetto.
- Campo di validazione, ad indicare se l'associazione tra i due record è da considerarsi valida (1) o meno (0); valori tra 0 e 1 vengono usati nei casi in cui non ci siano elementi sufficienti per capire quale sia il match valido, ma tali elementi siano sufficienti a restringere la scelta tra due o più possibilità.

Le tabelle di match concretizzano quindi una relazione molti-a-molti tra i due concetti, in cui ogni record specifica la distanza e la validità tra ogni coppia di elementi.

3.2.1 Match tra residenti e titolari

Il primo concetto comune individuato nello schema riconciliato è quello delle persone, memorizzate in anagrafe come residenti e in catasto come titolari. Dato il contesto, è ragionevole pensare che la maggior parte dei titolari di proprietà sul suolo cesenate siano residenti nello stesso comune; per questo motivo si è deciso di realizzare il match approssimato tra le due entità. Trattandosi di concetti simili, è logico che i campi delle rispettive tabelle siano anch'essi condivisi: nome, cognome, codice fiscale, sesso, data di nascita e codice del comune di nascita sono infatti presenti in entrambe le strutture, pertanto sono questi i campi di riferimento utilizzato.

Prima di procedere con l'implementazione del match è necessario capire quale sia la relazione tra le due entità. Da una parte si trova l'elenco dei residenti, il quale risulta qualitativamente buono e con pochi (e trascurabili) errori nei singoli record. Dall'altra parte, l'elenco dei titolari catastali presenta diversi problemi: oltre ai singoli errori di ortografia, può capitare che una stessa persona sia elencata più volte ma con dati leggermente diversi (il cognome scritto diversamente, la data di nascita errata, ecc.). Questo significa che per ogni persona in anagrafe potrebbero corrispondere uno o più titolari (o anche nessuno); al contrario, per ogni persona in catasto può corrispondere al massimo un unico residente. Per questo motivo, si

decide di impostare il match approssimato partendo dai titolari catastali e cercando, per ciascuno di essi, quale sia il migliore (ed unico) aggancio tra i residenti.

Per eseguire il confronto tra i record dei residenti e quelli dei titolari è necessario specificare il modo in cui deve essere calcolata la similarità tra i vari campi.

Tabella 1 - Calcolo della similarità tra i campi dei residenti e dei titolari

Campi confrontati	Regola utilizzata
Nome e cognome	Si tokenizzano le stringhe e si calcola la somma pesata delle distanze di Levenshtein normalizzate tra i singoli token.
Data di nascita	Si calcola la distanza di Levenshtein normalizzata, considerando le due date come semplici stringhe (i formati delle date sono ovviamente identici)
Sesso	Si considera 0 se c'è corrispondenza, 1 altrimenti.
Codice del Comune di nascita	Si considera 0 se c'è corrispondenza, 1 altrimenti.

In secondo luogo vengono definite le regole progressive per l'accettazione delle associazioni individuate:

Tabella 2 - Regole progressive per i match tra residenti e titolari

Regola	Descrizione
Regola 1	Match perfetti (uguaglianza su tutti i campi). Si è deciso di considerare perfetti anche i match in cui c'è corrispondenza sul solo codice fiscale (in quanto esso riassume i valori di tutti gli altri campi).
Regola 2	La distanza sul nome deve essere minore o uguale a 0.2 e le due data di nascita non devono essere nulle.
Regola 3	La distanza del nome deve essere minore o uguale a 0.3, le due data di nascita non devono essere nulle e deve esserci corrispondenza sul sesso e sul codice del comune di nascita.

La regola 1 è stata implementata con una semplice query SQL; le associazioni individuate vengono inserite nella tabella di match col campo di validazione già impostato a 1. Le regole 2 e 3 sono state invece implementate con uno script C#,

in cui vengono considerati i soli titolari per cui il match non è stato ancora trovato; a tal proposito viene precedentemente creata una tabella temporanea, *titolari_no_match*, per contenere i suddetti titolari non ancora agganciati. In questi casi, le associazioni individuate vengono inserite nella tabella di match col campo di validazione a 0: è infatti possibile che, per ogni titolare, siano state trovate associazioni accettabili con più di un residente. Si rende quindi necessario un ultimo step, in cui bisogna scegliere - per ogni titolare - l'associazione migliore tra quelle individuate. Tale operazione viene implementata con una Trasformazione PDI; il criterio di scelta dell'associazione migliore si basa sul calcolo di una distanza complessiva (descritto col seguente pseudocodice):

```
distanzaComplessiva = distanzaNomeCognome + distanzaDataNascita;
if ( distanzaCF == 1 ) then distanzaComplessiva += 0.01;
if ( distanzaSesso == 1 ) then distanzaComplessiva += 0.01;
if ( distanzaCodComuneNascita == 1 ) then distanzaComplessiva += 0.01;
```

Figura 4 - Pseudocodice per il calcolo della distanza complessiva tra residente e titolare

Per ogni titolare viene infine scelta l'associazione col residente con cui è minima la distanza complessiva.

A fronte dell'esecuzione del join approssimato, i risultati possono essere espressi col seguente grafico.

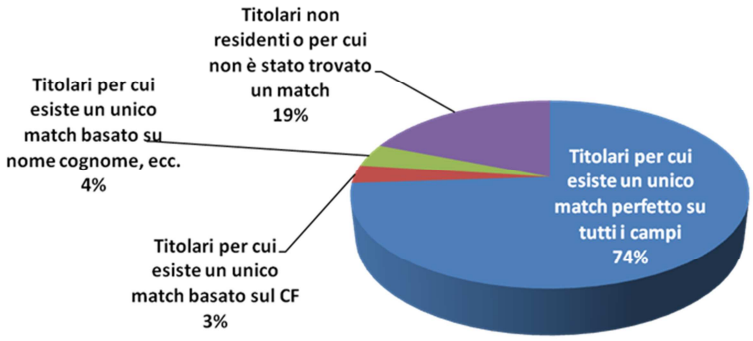

Figura 5 - Statistiche del match tra residenti e titolari

Le statistiche sono calcolate sul numero di titolari presenti in catasto, il quale ammonta ad 81212 persone. Come si può notare dal grafico, il join approssimato ha consentito un aumento complessivo di match pari al 7% (corrispondente a circa un quarto di quelli che erano prima i titolati non agganciati). Tuttavia, nonostante il join approssimato, il 19% dei titolari rimane scollegato dall'anagrafe.

Questo fenomeno può essere causato dalla presenza di troppi errori nei dati dei titolari, tali da rendere inefficaci le tecniche di join approssimato. Più probabilmente, però, la spiegazione risiede nel fatto che tali soggetti potrebbero non essere mai stati residenti nel Comune di Cesena dal 1999 (data di inizio dell'archivio anagrafico digitale) in avanti.

3.2.2 Match tra toponimi

Il secondo importante concetto comune individuato nello schema riconciliato è quello dei toponimi (ossia i nomi delle strade), presente (oltre che in toponomastica) negli indirizzi di residenza, negli indirizzi degli immobili e negli indirizzi delle utenze. In toponomastica, i nomi delle strade sono definiti in maniera univoca ed ufficiale; la tabella *t_strade* costituisce pertanto il punto di riferimento su cui basarsi. In anagrafe, gli indirizzi di residenza specificano il toponimo attraverso l'identificativo numerico (*cod_via*) utilizzato in toponomastica; in questo caso, quindi, l'integrazione è già esistente. Per quanto riguarda invece il catasto e le utenze dei consumi, i toponimi sono scritti secondo i più variegati formalismi; inoltre, una stessa strada può comparire in tanti modi diversi all'interno della stessa tabella. In questi contesti, la bonifica dei toponimi si rende sicuramente necessaria. Infine, siccome il contesto del catasto è identico a quello delle utenze, da qui in avanti si citerà unicamente il catasto (per pura semplicità), pur riferendosi ugualmente anche alle utenze.

Analogamente al join tra residenti e titolari, la relazione teorica tra le strade in toponomastica e le strade in catasto è una 1-N: ogni strada nel catasto corrisponde ad un'unica strada in toponomastica, ma - al contrario - una strada in toponomastica può trovare più corrispondenze tra le strade nel catasto. Per questo motivo, si decide di impostare il match approssimato partendo dai toponimi del catasto e cercando, per ciascuno di essi, quale sia il migliore (ed unico) aggancio tra i toponimi in toponomastica.

Prima di descrivere il meccanismo di match, è bene sottolineare come i join approssimati sui toponimi siano stati materializzati in quattro diverse tabelle di match - una per ogni tabella di partenza (indirizzi catastali e utenze luce, acqua e gas). Le tabelle di match non specificano però nessun identificativo delle suddette

tabelle, bensì contengono i soli campi descrittori delle strade; un possibile sviluppo futuro potrebbe quindi riguardare l'unificazione di queste tabelle di match, per gestire il join dei toponimi con un'unica tabella.

I campi su cui si basa il join approssimato sono dati dagli unici due descrittori delle strade, ossia le combinazioni di DUG e toponimo, perciò il calcolo della distanza tra due strade avviene con un unico confronto che considera entrambi i campi. Prima di iniziare tale confronto, sulla strada del catasto viene eseguita una bonifica iniziale: è stato infatti rilevato un frequente utilizzo errato del DUG qualora esso sia diverso dal classico "VIA". Più precisamente, invece di separare correttamente DUG e toponimo, il DUG viene incluso in forma estesa nel campo del toponimo, mentre nel campo del DUG viene indicato il generico "VIA". Per fare un esempio, invece della coppia {"VLO", "Sala"}, nelle tabelle viene spesso trovata la coppia {"VIA", "Vicolo Sala"}. In questi casi, il mero confronto delle due coppie darebbe un esito molto negativo: il DUG risulta infatti discordante e la presenza della parola "Vicolo" causa una distanza molto elevata tra i due toponimi. Per risolvere questo problema, si è deciso di operare un primo controllo di tutti i toponimi, rimuovendo da essi le parole chiave relative ai DUG (come "Vicolo", "Corso", "Galleria", ecc.); inoltre dalla parole chiave eventualmente rimossa viene ricavato il corrispettivo DUG. E' bene specificare che queste bonifiche vengono fatte soltanto a run-time e che non vengono in alcun modo memorizzate.

Entrando nel merito del confronto vero e proprio, l'algoritmo implementato risulta molto simile a quello utilizzato per il confronto del nome e del cognome nel join tra residenti e titolari: si tokenizzano le stringhe dei toponimi e si calcola la somma pesata delle distanze di Levenshtein normalizzate tra i singoli token. Oltre alla distanza totale tra i due toponimi, si è deciso di tener conto anche della distanza minima tra due token; in questo modo si ha a disposizione un parametro in più per decidere quali siano i match migliori. Un ultimo passaggio riguarda infine il confronto dei DUG: il DUG della strada in toponomastica viene confrontato sia col DUG della strada in catasto, sia col DUG eventualmente estratto dal toponimo di quest'ultima. Se non si trova corrispondenza, la distanza totale viene incrementata di 0.05.

Le regole progressive per l'accettazione delle associazioni individuate vengono di seguito definite:

Tabella 3 - Regole progressive per i match tra toponimi

Regola	Descrizione
Regola 1	Match perfetti (il DUG e il toponimo coincidono).
Regola 2	(distanzaTotale ≤ 0.35 AND distanzaMinima == 0) OR (distanzaTotale ≤ 0.30 AND distanzaMinima > 0)

La regola 1 è stata implementata con una semplice query SQL; le associazioni individuate vengono inserite nella tabella di match col campo di validazione già impostato a 1. La regola 2 è stata invece implementata con uno script C#, in cui vengono considerate le sole strade del catasto per cui il match non è stato ancora trovato; a tal proposito viene precedentemente creata una tabella temporanea per contenere le suddette strade non ancora agganciate. In questi casi, le associazioni individuate vengono inserite nella tabella di match col campo di validazione impostato a 0: è infatti possibile che, per ogni strada in catasto, siano state trovate associazioni accettabili con più di una strada in toponomastica. Si rende quindi necessario un ultimo step, in cui bisogna scegliere - per ogni strada in catasto - l'associazione migliore tra quelle individuate. Tale operazione viene implementata con una Trasformazione PDI, la quale permette di eseguire in pochi passi le seguenti scelte.

- Per ogni strada del catasto si sceglie come valida la strada in toponomastica per cui la distanza totale è più bassa.
- In caso di pareggio tra due o più strade in toponomastica, si sceglie come valida quella per cui la distanza minima è più bassa.
- In caso di ulteriore pareggio, non esistono altri parametri di decisione della scelta migliore. In questa eventualità, a ciascuna delle *n* strade toponomastiche rimaste viene assegnato un valore di validità pari ad 1/n. In questo modo, seppur non si sappia quale sia il match corretto, resta noto il fatto che il match corretto è - con tutta probabilità - uno di quelli indicati. Per fare un esempio concreto, la strada {"VIA", "Emi-

lia"} avrà uguali distanze dalle strade {"VIA", "Emilia Ponente"} e {"VIA", "Emilia Levante"}; in questo caso, nella tabella di match verranno aggiunte entrambe le associazioni con un valore di validità uguale a 0.5 (il significato è: "non si sa quale delle due sia, ma di sicuro è una di quelle due").

A fronte dell'esecuzione del join approssimato, i risultati possono essere espressi col seguente grafico; i dati riportati riguardano i match a partire dalle utenze dei consumi di acqua, ma le statistiche sono pressoché identiche per tutti e quattro i match di questo tipo.

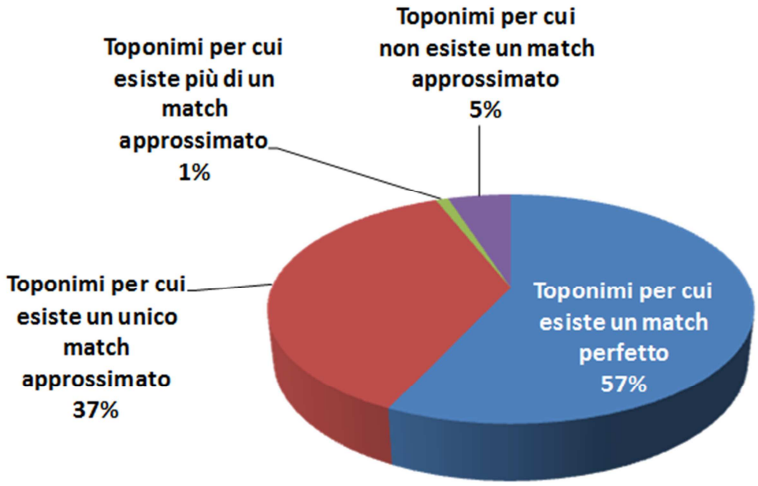

Figura 6 - Statistiche sul match tra i toponimi nelle utenze dei consumi d'acqua e quelli in toponomastica

Innanzitutto, dalle statistiche si può notare quanto siano bassi i match perfetti sui toponimi, specialmente se si considera il paragone con i match perfetti sulle persone: mentre le differenze tra nomi e cognomi riguardano principalmente errori manuali, i toponimi possono essere scritti in tanti modi diversi, per cui senza l'utilizzo dei formalismi ufficiali ci si ritrova con una percentuale bassa di agganci perfetti. Nella Figura 7 si può notare la varietà di formalismi che possono essere utilizzati in determinati casi: la prima riga di entrambe le tabelle indica il toponimo ufficiale, mentre i numeri indicano la quantità di indirizzi catastali che utilizzano il rispettivo toponimo.

SAVIO IN S.ANDREA	5
DEL SAVIO S.ANDREA IN BAGNOLO	3
SAVIO (LOC. S.ANDREA IN BAGNOLO)	8
SAVIO - S. ANDREA IN B.	10
SAVIO - S. ANDREA IN BAGNOLO	6
SAVIO - S.ANDREA IN BAGNOLO	36
SAVIO IN S. ANDREA	80
SAVIO IN SAN ANDREA	4
SAVIO IN SANT` ANDREA	50
SAVIO IN SANT`ANDREA	61
SAVIO S. ANDREA IN BAGNOLO	4
SAVIO S.ANDREA	2
SAVIO S.ANDREA IN BAGNOLO	1
SAVIO-S.ANDREA IN B.	3
VIC.SAVIO S ANDREA IN B	1

CHIESA DI S.ANDREA	0
CHIESA DI S. ANDREA	35
CHIESA DI S. ANDREA IN BAGNOLO	2
CHIESA DI S.ANDREA IN BAGNOLO	10
CHIESA DI SANT` ANDREA	30
CHIESA DI SANT`ANDREA IN BAGNOLO	1
CHIESA DI SANT``ANDREA IN BAGNOLO	27
CHIESA S.ANDREA IN B.	1

Figura 7 - Esempi di toponimi utilizzati negli indirizzi catastali

In secondo luogo, le stesse statistiche evidenziano l'efficacia delle tecniche di join approssimato, grazie alle quali si riesce a raggiungere la quasi totalità degli agganci. Il fenomeno si nota maggiormente sulle utenze, dove gli indirizzi impiegano formalismi diversi ma non presentano quasi mai degli errori. Nel catasto, invece, la qualità complessivamente minore degli indirizzi causa una percentuale di agganci leggermente minore (87% di agganci totali, partendo dal 55% dei match perfetti).

3.2.3 Match tra coordinate catastali e indirizzi di residenza

Il match tra le strade ha permesso di risolvere il problema della bonifica dei nomi delle strade in catasto, creando un "ponte di collegamento" tra le UIU e gli indirizzi di residenza che non passi dalla relazione titolari-residenti. Tuttavia, tale collegamento risulta incompleto a causa dell'assenza della numerazione civica interna negli indirizzi in catasto. Questo significa che sapere "chi abita in un determinato immobile" o "chi possiede l'immobile in cui abita una determinata persona" risulta possibile solo per case singole, in cui non sono presenti più appartamenti con civici interni diversi.

L'idea per questo tipo di match nasce quindi dall'esigenza di sopperire alla mancanza del civico interno negli indirizzi delle UIU, cercando una via alternativa

che completi il collegamento. Diversamente dai match precedentemente implementati, però, quello tra coordinate catastali e indirizzi di residenza si basa su un concetto comune più "astratto", tale per cui le classiche tecniche di fusione approssimata non sono applicabili. Il concetto di fondo consiste nel fatto che ad ogni coordinata catastale corrisponde univocamente un indirizzo di residenza (completo di numerazione civica esterna ed interna). Le due informazioni costituiscono infatti due formalismi diversi per rappresentare la stessa cosa: un'identificazione geografica di una UIU. Sebbene il match tra questi due concetti non sia possibile dal punto di vista prettamente strutturale, la correlazione tra di essi è sicuramente esistente.

Non potendo confrontare direttamente i concetti da unire, ciò che si vuol fare è cercare delle condizioni che permettano di stabilire l'associazione tra una coordinata catastale e un indirizzo di residenza. Le condizioni da cercare si concentrano sulle proprietà catastali di un residente: ciò che si ipotizza è che, se un residente possiede una UIU di tipo residenziale il cui indirizzo è compatibile col suo indirizzo di residenza, l'immobile di residenza corrisponderà probabilmente a tale UIU. La situazione è chiaramente più complessa se le proprietà sullo stesso indirizzo sono più di una, perciò si è deciso di applicare una serie di regole progressive per individuare le associazioni:

Tabella 4 - Regole progressive per i match tra coordinate catastali e indirizzi di residenza

Regola	Descrizione
Regola 1	Il residente possiede una sola UIU di tipo residenziale ed essa si trova nella strada in cui risiede.
Regola 2	Il residente possiede due o più UIU di tipo residenziale, ma solo una di esse si trova nella strada in cui risiede.
Regola 3	Il residente possiede due o più UIU di tipo residenziale che si trovano nella strada in cui risiede, ma solo una di esse si trova nello stesso civico esterno.
Regola 4	Il residente possiede due o più UIU di tipo residenziale che si trovano nella strada e nel civico in cui risiede.

Innanzitutto, è bene specificare che - diversamente rispetto agli altri match - alla progressività delle regole non corrisponde un progressivo aumento delle tolleranze di accettazione dei match; la numerazione indica solamente l'ordine di applicazione delle regole. In secondo luogo, le associazioni individuate tramite le regole 1, 2 e 3 vengono inserite nella tabella di match col campo di validazione impostato a 1; per la regola 4, invece, il campo di validazione è posto uguale a 1/*n* (dove *n* indica il numero di UIU rimaste a quel punto della selezione). Il meccanismo fin qui descritto è stato implementato con un'unica e complessa Trasformazione PDI.

Entrando nel merito dei risultati, va specificato che questo match si basa su una condizione molto flebile, in cui non c'è un grado di certezza assoluta e in cui si perdono sicuramente tutti i casi in cui un'UIU non è abitata da uno dei proprietari. Gli effetti di questa considerazione sono evidenti nei risultati:

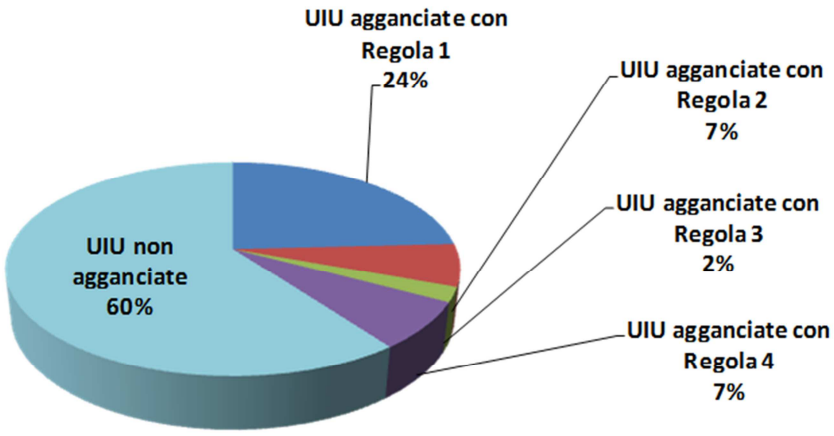

Figura 8 - Statistiche sul match tra coordinate catastali e indirizzi di residenza

Se si considera che gli agganci con Regola 4 sono soltanto approssimativi (in quanto espressione di un dubbio tra più indirizzi di residenza che differiscono nel solo civico interno), soltanto il 31% delle UIU residenziali risultano associate ad un indirizzo di residenza. Tuttavia, nonostante le percentuali siano molto basse, gli agganci trovati non sono comunque da scartare: anche se pochi, essi forniscono un'informazione precedentemente sconosciuta ed utile. Infatti, tali informazioni vengono poi utilizzate nei pattern (vedi capitolo 4), quando - ad esempio - si vuole cercare l'associazione tra un'utenza ed una UIU.

3.3 Estensione del database

Nella prima fase del progetto, il database riconciliato è stato costruito integrando le banche dati del Comune (anagrafe e toponomastica) e alcuni di quelle fornite dall'Agenzia delle Entrate (catasto e utenze dei consumi). Tuttavia, al fine di individuare dei complessi pattern di evasione è stato necessario arricchire il database con ulteriori informazioni. In questo paragrafo vengono pertanto descritte le varie banche dati che sono state successivamente aggiunte. Inoltre, per ogni integrazione al database viene fornito il diagramma ER focalizzato sulle singole modifiche: con colori chiari vengono indicate le entità esistenti e con colori scuri le integrazioni effettuate.

3.3.1 Indirizzi di residenza

Nella prima fase del progetto, dalla banca dati dell'anagrafe sono già stati prelevati gli indirizzi di residenza; non si tratta però di informazioni storicizzate: come spiegato nel paragrafo 3.1.1, quella importata è solamente la fotografia delle residenze al 31 dicembre 2010. Tuttavia, per applicare i pattern di evasione è necessario conoscere lo storico completo delle residenze, in modo da verificare gli eventuali spostamenti delle persone (in termini di residenze) e da controllare le situazioni in diversi momenti temporali. Nella banca dati dell'anagrafe, lo storico completo degli indirizzi di residenza risulta correttamente memorizzato, pertanto si è deciso di sostituire la tabella degli indirizzi di residenza precedentemente importata (*a_indirizzi*) con una nuova tabella in grado di ospitare le informazioni storicizzate. Tale tabella è stata chiamata *at_residenti_indirizzi*, in quanto - sebbene sia memorizzata nella banca dati anagrafica - essa materializza in pratica l'associazione molti-a-molti tra i residenti (*a_residenti*) e gli indirizzi toponomastici (*t_indirizzi*).

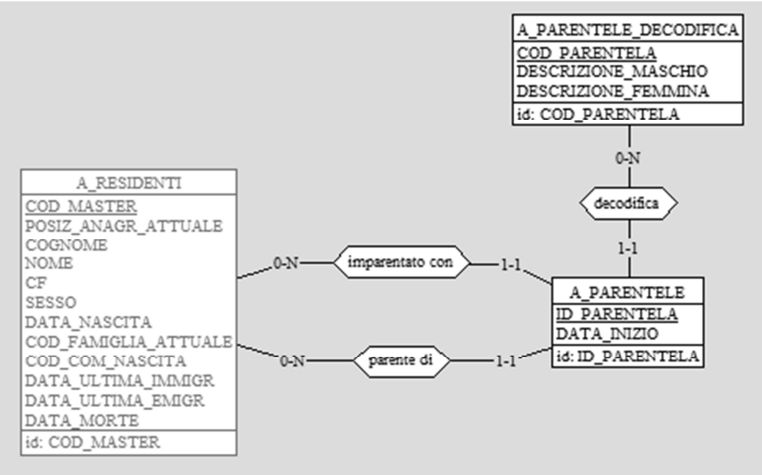

Figura 9 - Diagramma ER focalizzato sugli indirizzi di residenza

L'importazione dei dati dal database dell'anagrafe al database riconciliato è stata impostata con una trasformazione PDI, i cui step possono essere così riassunti:

- Caricamento di tutti i record dallo storico dell'anagrafe.
- Esclusione dei record in cui mancano il codice della via e il numero civico esterno (fondamentali per l'aggancio a *t_indirizzi*).
- Bonifica del civico-bis e dell'interno: se sono nulli, vengono sostituiti con "-1".
- Scrittura dei dati in *at_residenti_indirizzi*.

In seguito al caricamento dei dati, alcune query SQL sono state eseguite per bonificare i dati, partendo dall'aggiunta di un campo auto-incrementate come chiave primaria della tabella e dall'eliminazione dei record in cui la data di fine validità era minore della data di inizio. In secondo luogo si è dovuto risolvere il problema degli agganci degli indirizzi a *t_indirizzi*: infatti, mentre per i residenti è presente il *cod_master* (la chiave primaria in *a_residenti*) come chiave esterna, gli indirizzi sono indicati come combinazione di {*cod_via*, *civico, civico_bis, interno*}. Sebbene venga quindi impiegato il formalismo corretto per le strade, l'assenza di un vincolo referenziale può causare problemi di integrità (oltre a rendere più complesse le operazioni di join). Sono state quindi eseguite una serie di query SQL con lo scopo ultimo di implementare il vincolo referenziale, aggiungendo in *at_indirizzi_residenti* l'identificativo dell'indirizzo come chiave esterna.

- Bonifica di *t_indirizzi* per eliminare i record con *cod_via* e *civico* nulli ed eliminare record duplicati.
- Bonifica della numerazione civica interna in *t_indirizzi*: mentre in a*t_residenti_indirizzi* esiste un unico campo (*interno*), in *t_ indirizzi* la stessa informazione è separata in due campi (*interno* e i*nterno_bis*). Inoltre, non esiste un formalismo standard per la memorizzazione dell'interno-bis, per cui - in entrambe le tabelle - l'interno-bis può essere indicato con o senza "/" (ad esempio, in *at_indirizzi_residenti* si può trovare sia {"8A"} che {"8/A"}, mentre in *t_indirizzi* si può trovare sia {"8"; "A"} che {"8"; "/A"}). Si è pertanto deciso di creare due nuovi campi in *t_indirizzi* (*interno_edit_1* e *interno_edit_2*) in cui unire i campi *interno* e *interno_bis* utilizzando rispettivamente il formalismo con e senza "/".
- Confronto dei campi di *at_residenti_indirizzi* con quelli di *t_indirizzi* per importare nella prima tabella l'identificativo degli indirizzi.

Al termine di questa procedura, non tutti gli indirizzi hanno trovato un riscontro diretto in *t_indirizzi*; circa il 14% dei record in *at_residenti_indirizzi* non ha infatti trovato un aggancio. Ad esempio, ci sono indirizzi di residenza che specificano un civico interno non elencato in toponomastica; al contrario, ci sono anche indirizzi di residenza che non specificano l'interno, quando invece quest'ultimo è richiesto in toponomastica. In questi casi, il campo contenente l'identificativo dell'indirizzo è stato lasciato nullo. Tuttavia, solo 2% degli indirizzi di residenza non agganciati sono ancora validi; ciò significa che quelli non agganciati sono in gran parte indirizzi vecchi e potenzialmente meno interessanti. Se si restringe il conto ai soli indirizzi di residenza tuttora validi, la percentuale di agganci in toponomastica è pari al 99,5%.

3.3.2 Relazioni di parentela

Un'altra informazione utile estraibile dall'anagrafe è data dalle relazioni di parentela tra i residenti. Questi dati risultano di grande importanza nell'analisi dei pattern: ad esempio, sapere se una persona è residente in una casa posseduta da parenti o da estranei è fondamentale per avanzare l'ipotesi che tale persona sia o meno in affitto.

Prima di descrivere il procedimento di integrazione, va citato il fatto che i dati sulle relazioni di parentela non sono perfetti: le informazioni originali (dalle quali essi sono stati manualmente ricavati) presentano infatti alcuni errori strutturali, per cui - oltre al fatto che gli identificativi delle persone non corrispondono ai rispettivi *cod_master* in *a_residenti* - una stessa persona può comparire più volte con identificativo diverso. Per questo motivo, la trasposizione delle parentele sulla tabella dei residenti - seppur sia stata effettuata con le dovute attenzioni - potrebbe presentare alcuni errori. Tuttavia, alla conclusione del progetto non è ancora stato rilevato nessun errore.

Un'ultima osservazione deve essere infine fatta sui matrimoni, la cui gestione è sicuramente più complessa: diversamente rispetto ai legami di sangue (i quali sono chiaramente immutabili dalla nascita), i matrimoni possono essere seguiti da divorzi e da matrimoni successivi. A causa di alcune difficoltà nella ricostruzione dello storico dei matrimoni, nelle relazioni di parentela sono stati inclusi i soli matrimoni in essere. Ciò non ha costituito un problema per il progetto, dal momento che per le indagini svolte era necessario conoscere i soli matrimoni attualmente validi.

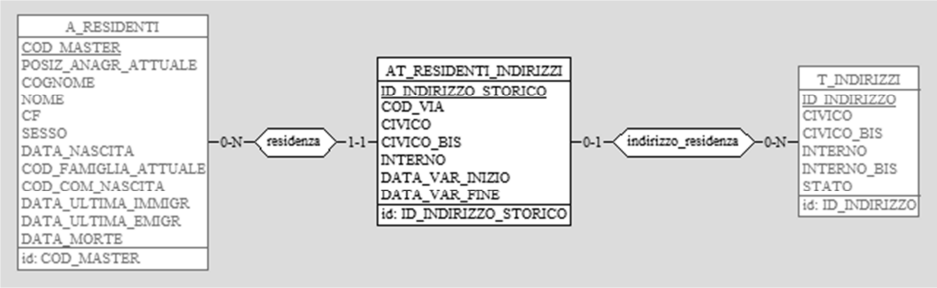

Figura 10 - Diagramma ER focalizzato sulle parentele

L'aggiunta delle relazioni di parentela ha richiesto la creazione di due tabelle, *a_parentele_decodifica* (in cui viene fornita la descrizione testuale dei codici delle parentele) e a_parentele (in cui vengono memorizzate le varie relazioni). In particolare, le relazioni di parentela sono memorizzate come associazioni tra due cod_master; inoltre, ogni relazione risulta ripetuta con i soggetti invertiti: se a e b sono due residenti - padre e figlio -, in a_parentele esisterà un record per specificare "a padre di b" e un altro record per specificare "b figlio di a". Il processo di importazione di entrambe le tabelle è stato implementato con due semplici Tra-

sformazioni PDI, alle quali non hanno fatto seguito altre operazioni di bonifica o di integrazione.

3.3.3 Contratti d'affitto

Avendo deciso di incentrare il progetto sulla lotta agli affitti in nero, la banca dati dei contratti d'affitto costituisce una fonte di assoluta importanza. Quando si individua una situazione sospetta, sapere se il contratto d'affitto è stato o meno stipulato è ovviamente necessario per confermare il sospetto o per scartare il caso. Come anticipato nel paragrafo 1.3.3, tuttavia, i dati (inviati al Comune dall'Agenzia delle Entrate e relativi agli anni 2009 e 2010) presentano delle forti limitazioni:

- Il 75% dei contratti d'affitto relativi a immobili non specifica l'immobile su cui il contratto è stato stipulato.
- Del restante 25% (in cui i dati degli immobili sono specificati), la metà dei contratti riguardano immobili che non si trovano nel Comune di Cesena. Il motivo risiede nel fatto che un contratto può essere depositato in qualunque comune (non necessariamente quello in cui si trova l'immobile), unito anche al fatto che il Comune ha accesso ai soli contratti depositati all'ufficio di Cesena; ne consegue che:
    - La metà dei contratti disponibili riguardano immobili relativi ad altri comuni.
    - Di riflesso, è stimabile che la metà dei contratti relativi ad immobili di Cesena siano stati depositati negli uffici di altri comuni.

A causa di questi problemi (che non sono chiaramente risolvibili dal punto di vista informatico), il potere informativo di questa banca dati cala drasticamente. Tuttavia, il problema non è bloccante per il progetto: i contratti d'affitto avrebbero avuto il ruolo di dare una prima conferma delle situazioni sospette, ma l'individuazione stessa delle situazioni sospette si basa su parametri diversi. La decisione è stata quindi quella di utilizzare i pochi contratti disponibili per escludere eventuali falsi positivi rilevati.

I contratti d'affitto ricevuti si suddividono in 3 tipologie:

- “Atti relativi a contratti di locazione registrati in ufficio”.
- “Contratti di locazione registrati in ufficio”.
- “Contratti di locazione pervenuti telematicamente”.

Il primo formato citato è quello utilizzato fino al 2009; tale formato non prevedeva in alcun modo la memorizzazione degli immobili su cui avveniva la locazione. Dal 2010 sono entrati in vigore gli altri due formati, che permettono la memorizzazione degli immobili e che distinguono i contratti registrati in ufficio da quello compilati per via telematica; sebbene i dati siano concettualmente gli stessi, i due formati presentano delle leggere differenze nella memorizzazione dei dati. In quanto ai contenuti, è stato rilevato che i contratti registrati telematicamente sono qualitativamente migliori, probabilmente grazie al supporto e ai controlli implementati nell’interfaccia telematica; in particolare, i dati sull’immobile affittato sono sempre presenti. Nella registrazione in ufficio, invece, i dati sull’immobile vengono quasi sempre tralasciati; ciò è probabilmente dovuto all’assenza di controlli o alla bontà degli addetti nell’accettare contratti che non specificano tutti i dati. Sfortunatamente, il rapporto tra i contratti registrati telematicamente e quelli registrati in ufficio è di circa 1:9.

I contratti d’affitto sono stati consegnati dall’Agenzia delle Entrate sotto forma di tracciati testuali: uno per gli atti del 2009, uno per i contratti registrati in ufficio nel 2010 e uno per i contratti registrati telematicamente nel 2010. Per ciascuno di essi è stato anche fornito il manuale per l’interpretazione dei dati. La decodifica di questi tracciati e la loro importazione nel database riconciliato si è suddivisa in due fasi: i dati sono stati innanzitutto importati in una banca dati del Sistema Informativo del Comune; successivamente, gli stessi dati sono stati poi importati nel database riconciliato.

I tracciati testuali presentano le informazioni in maniera denormalizzata; in particolare, il formato utilizzato per i record può essere riassunto nel modo seguente:

Tabella 5 - Struttura di base dei tracciati relativi ai contratti d'additto

Attributo	Descrizione
Tipo elemento	"A" indica un contratto, "B" indica un soggetto, "I" indica un immobile.
ID contratto	Una sequenza di informazioni (ufficio di registrazione, anno, serie, numero, sottonumero e progressivo) che, nel complesso, identificano un singolo contratto.
Info elemento	Una serie di informazioni che cambiano a seconda del tipo di elemento (primo campo).

Per l'importazione dei dati nel SI è stata utilizzata la funzionalità di Oracle Toad che prevede specificamente l'importazione di dati da un tracciato ad una tabella, consentendo tra l'altro di spezzare le righe in diversi punti e di memorizzare i vari "pezzi" in appositi campi. Dall'analisi del tracciato è evidente la necessità di creare tre distinte tabelle (una per ciascuna tipologia di elemento), ma la funzionalità utilizzata non permette di farlo direttamente. Pertanto, si è deciso di fare inizialmente uno "scarico", ossia di importare tutti i dati in un'unica tabella temporanea i cui campi siano generici e applicabili a tutti i tipi di elementi; tali campi sono gli stessi indicati poco sopra: *tipo_elemento*, *id_contratto* e *info_elemento*. Dopo aver eseguito lo scarico, sono state create delle query SQL per selezionare i dati dalla tabella temporanea e trasferirli (in base al tipo di elemento) nelle apposite tabelle. Al termine delle operazioni sono quindi state create e popolate le tabelle *con_contratti*, *con_soggetti* e *con_immobili* (in conformità alla sintassi usata nel SI).

A causa della struttura dei tracciati, le tabelle dei soggetti e degli immobili risultano ancora denormalizzate: se un soggetto o un immobile sono presenti in più contratti, le loro informazioni risultano infatti ripetute per ogni contratto (l'ID dei soggetti e degli immobili è composto dall'ID del contratto e da un numero progressivo). Questo significa che le relazioni tra la tabella dei contratti e quelle dei soggetti e degli immobili sono delle 1-N (ad ogni contratto possono corrispondere più soggetti/immobili, ma un soggetto/immobile è relativo ad un unico contratto). Per risolvere la situazione, una possibilità sarebbe quella di modificare lo schema, introducendo delle relazioni N-N ed eliminando i duplicati nelle tabelle

dei soggetti e degli immobili. Tuttavia, dal momento in cui l'utilizzo previsto di questi dati è molto limitato, si è deciso di lasciare immutata questa situazione.

Pur lasciandoli denormalizzati, i dati hanno comunque richiesto una serie di operazioni di bonifica, per correggere alcuni errori e per rendere più chiare alcune informazioni:

- Eliminazione di contratti e soggetti duplicati: data la possibilità per un singolo contratto di regolare l'affitto di più immobili contemporaneamente, ad ogni contratto possono essere associati più immobili; tuttavia, nei tracciati testuali non possono essere inseriti più di tre immobili (per probabili limitazioni nel software utilizzato dall'Agenzia delle Entrate). Se un contratto riguarda quindi quattro o più immobili, le informazioni di tale contratto (e dei rispettivi soggetti coinvolti) vengono ripetute per consentire l'inserimento di tutti gli immobili.
- Eliminazione di soggetti ripetuti più volte all'interno dello stesso contratto (in situazioni indipendenti da quella sopracitata).
- Correzione di date errate.
- Conversione in Euro di alcuni importi ancora in Lire.

Terminate le bonifiche, i dati sono stati importati nel database riconciliato con tre trasformazioni PDI. Il diagramma ER in Figura 11 mostra in che modo le tabelle relative ai contratti d'affitto si inseriscono nel database.

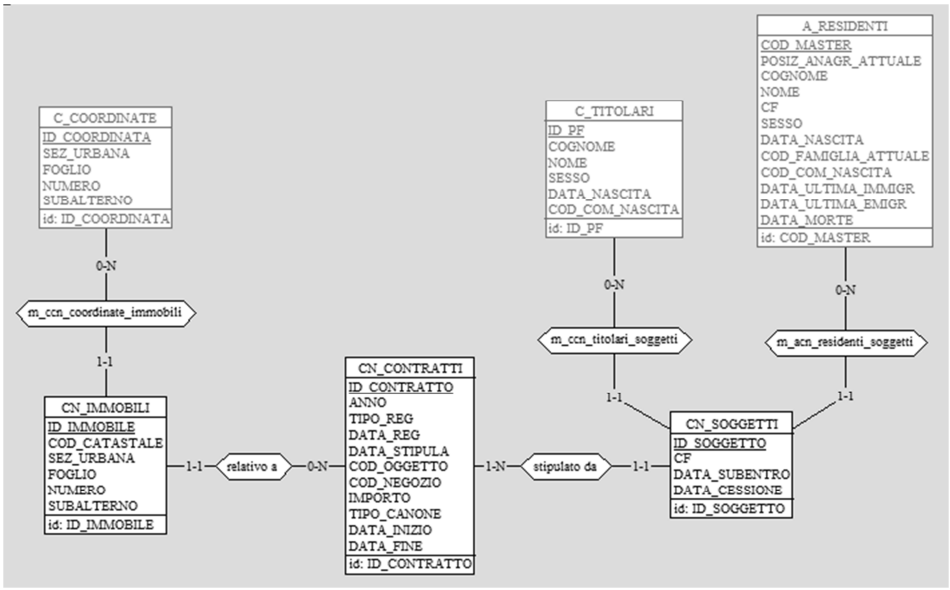

Figura 11 - Diagramma ER focalizzato sui contratti d'affitto

L'ultimo passo da eseguire riguarda infine l'integrazione vera e propria dei dati. Nei contratti d'affitto esistono infatti due concetti, i soggetti e gli immobili, che possono essere collegati a tabelle già esistenti nel database riconciliato: i residenti dall'anagrafe, i titolari e gli immobili dal catasto. Le informazioni a disposizione sui soggetti e sugli immobili sono però soltanto il minimo indispensabile: seppur siano previsti tutti i campi descrittivi, solamente il codice fiscale dei soggetti risulta compilato; analogamente, per gli immobili sono compilate le sole coordinate catastali. In questa situazione, è evidente l'assenza delle condizioni necessarie per implementare un join approssimato; si potrebbe studiare un algoritmo per il confronto intelligente dei codici fiscali e delle coordinate catastali, ma i risultati eventualmente ottenuti non giustificherebbero la quantità di tempo spesa per l'implementazione. Per questo motivo si è deciso di individuare solamente i match perfetti; in particolare, le tabelle di match implementate sono le seguenti.

Tabella 6 - Descrizione delle tabelle di match principali collegate ai contratti d'affitto

Tabella	Descrizione
m_acn_ residenti_soggetti	Implementa il collegamento tra i residenti (*a_residenti*) e i soggetti dei contratti (*cn_soggetti*); il match è stato trovato per il 41% dei soggetti.

m_ccn_ titolari_soggetti	Implementa il collegamento tra i titolari (*c_titolari*) e i soggetti dei contratti (*cn_soggetti*); il match è stato trovato per il 32% dei soggetti.
m_ccn_ coordinate_ immobili	Implementa il collegamento tra gli immobili in catasto (*c_coordinate*) e gli immobili dei contratti (*cn_immobili*). Il match è stato trovato per il 37% degli immobili; se si considerano però i soli immobili di Cesena relativi a contratti di tipo residenziale, i match aumentano al 96%.

Data la suddivisione dei soggetti in affittuari e locatari, sono state fatte le seguenti osservazioni: se un soggetto è affittuario, il collegamento più interessante è quello con i residenti piuttosto che con i titolari; dall'altra parte, se un soggetto è locatario, il collegamento più interessante è quello con i titolari piuttosto che con i residenti. Per questo motivo si è deciso di implementare altre due tabelle di match (omesse, per semplicità, dal diagramma in Figura 11):

Tabella 7 - Descrizione delle tabelle di match secondarie collegate ai contratti d'affitto

Tabella	Descrizione
m_acn_ residenti_affittuari	Implementa il collegamento tra i residenti (*a_residenti*) e i soggetti dei contratti (*cn_soggetti*) che compaiono come affittuari; si riporta che, rispetto al totale di affittuari agganciati, il 98% di essi rientra in questa tabella.
m_ccn_ titolari_locatari	Implementa il collegamento tra i titolari (*c_titolari*) e i soggetti dei contratti (*cn_soggetti*) che compaiono come locatari; si riporta che, rispetto al totale di locatari agganciati, il 96% di essi rientra in questa tabella.

In conclusione, il flusso di caricamento dei dati dei contratti d'affitto può essere riassunto nei seguenti step:

- Utilizzo della funzionalità di Oracle Toad per "scaricare" i dati dei tracciati in tabelle del SI.
- Query SQL per separare i dati in base alla tipologia degli elementi (contratti, soggetti e immobili) e per applicare una serie di bonifiche.
- Trasformazioni PDI per importare i dati dal SI al database riconciliato.
- Query SQL per implementare i match sui soggetti e sugli immobili.

3.3.4 Dichiarazioni dei redditi

Oltre ai contratti d'affitto, l'Agenzia delle Entrate ha fornito al Comune di Cesena anche alcuni dati sulle dichiarazioni dei redditi. Sebbene si tratti di informazioni solamente riassuntive, esse sono sicuramente utili per individuare i pattern di evasione. I dati delle dichiarazioni risultano già integrati con l'anagrafe, perciò la loro importazione si concretizza con un'unica e semplice Trasformazioni PDI.

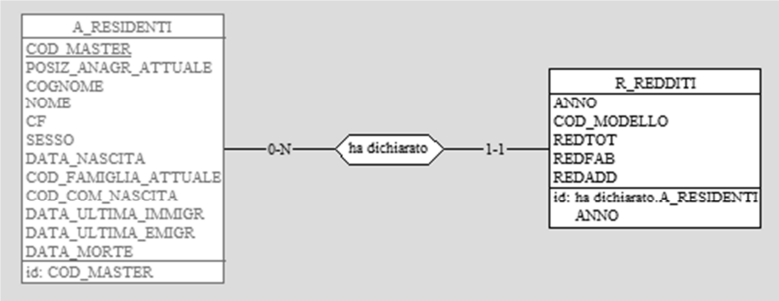

Figura 12 - Diagramma ER focalizzato sulle dichiarazioni dei redditi

Le informazioni importate includono il tipo di dichiarazione (*cod_modello*), il reddito totale (*redtot*), il reddito da fabbricati (*redfab*) e il reddito imponibile (*redadd*). Il principale scopo previsto per questi dati consiste nel confrontare il patrimonio economico di un soggetto con il suo patrimonio immobiliare e verificarne la congruenza. A tal proposito, alle informazioni riassuntive dei redditi di un soggetto si è deciso di affiancare una serie di informazioni riassuntive sulle sue proprietà catastali. Tali informazioni comprendono:

- Il numero di fabbricati (di tipo residenziale e non) e di terreni posseduti.
- La rendita catastale totale delle proprietà del soggetto.
- Il rapporto tra la rendita catastale totale e il reddito totale.
- Il rapporto tra la rendita catastale totale e il reddito da fabbricati.
- Il numero di utenze intestate per ogni tipo di consumo.
- Il numero di fabbricati posseduti le cui titolarità sono iniziate in corso d'anno o che presentano errori di incongruenza sulle quote di possesso.

Questi campi calcolati sono stati aggiunti alla stessa tabella con le dichiarazioni dei redditi. Il popolamento dei suddetti è stato però effettuato per il solo 2010,

ovvero l'anno in cui è previsto il concentramento delle indagini sugli affitti in nero. Per le operazioni di calcolo è stato necessario creare una nuova tabella, *c_uiu_flag*, in cui memorizzare informazioni utili sugli immobili in determinati momenti temporali, ovvero:

- L'anno a cui si riferiscono i dati.
- Un flag che indica se nell'immobile è residente un proprietario, un parente di proprietario o altre persone.
- Un flag che indica se, nel corso dell'anno, il totale delle quote di possesso dell'immobile corrisponde al 100%.
- Un flag che indica se, nel corso dell'anno, il totale delle quote di usufrutto coincide col totale delle quote di nuda proprietà.
- Il numero medio di occupanti nel corso dell'anno, stimato in base ai consumi eventualmente associati all'immobile.
- Il grado di affidabilità della stima del numero di occupanti.

Come per *r_redditi*, i campi sopracitati sono stati calcolati (attraverso diverse query SQL) per la sola annata del 2010.

In conclusione, il processo di caricamento dei dati può essere riassunto nei seguenti step:

- Trasformazione PDI per importare i dati delle dichiarazioni dei redditi.
- Serie di query SQL per la creazione e il popolamento di *c_uiu_flag*.
- Serie di query SQL per terminare il popolamento di *r_redditi*.

3.3.5 Censimento

A circa un anno di distanza dal censimento (avvenuto il 9 ottobre 2011), l'ISTAT ha inviato al Comune di Cesena le informazioni raccolte sul rispettivo territorio. I dati rilevanti per questo progetto riguardano i domicili dichiarati dalle persone, le quali sono state suddivise in tre categorie:

- Residenti censiti: persone che hanno dichiarato (nel censimento) di abitare a Cesena e che in anagrafe risultano effettivamente residenti a Cesena.

- Residenti non censiti: persone residenti a Cesena di cui non risulta il censimento (o che hanno dichiarato di risiedere in altri comuni).
- Censiti non residenti: persone che hanno dichiarato (nel censimento) di abitare a Cesena, ma che in anagrafe non risultano residenti a Cesena.

La categoria di dati più interessanti è quella dei "censiti non residenti": essi sono infatti dei potenziali affittuari su cui effettuare dei controlli.

I dati originali sono stati consegnati dall'ISTAT sotto forma di file Excel. Il caricamento nel database riconciliato è stato quindi eseguito attraverso delle query SQL, utilizzando il formalismo specifico di MySQL per la lettura dei dati da un file Excel. Successivamente sono state apportate alcune bonifiche, come la correzione di alcuni DUG e l'utilizzo dei formalismi standard per la memorizzazione del sesso e della data di nascita delle persone.

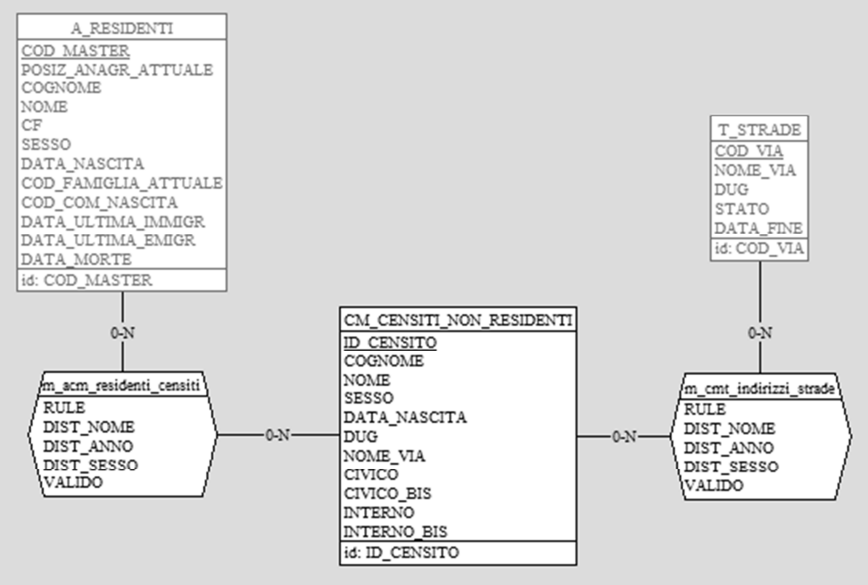

Figura 13 - Diagramma ER focalizzato sui dati del censimento

Al termine del caricamento, i dati sono stati poi integrati nel database riconciliato; si è resa quindi necessaria l'applicazione delle tecniche di join approssimato sui concetti comuni individuati: le persone (da agganciare ai residenti) e le strade (da agganciare a quelle in toponomastica). In entrambi i casi, le procedure di join sono identiche a quelle utilizzate per i join approssimati effettuati in precedenza,

pertanto si rimanda ai paragrafi 3.2.1 e 3.2.2 per la spiegazione dei meccanismi. Gli step previsti per l'esecuzione del join approssimato sono i seguenti:

1. Query SQL per individuare i match perfetti. I match individuati sono automaticamente validi.
2. Query SQL per creare le tabelle temporanee delle persone e delle strade per cui non è ancora stato trovato un match.
3. Script C# per individuare i match approssimati. I match individuati restano non validati.
4. Trasformazione PDI per validare il migliore tra i match approssimati individuati per ogni persona e per ogni strada.

3.3.6 Schema dati finale

Al termine dell'integrazione dei nuovi dati, il database riconciliato presenta l'architettura indicata in Figura 14.

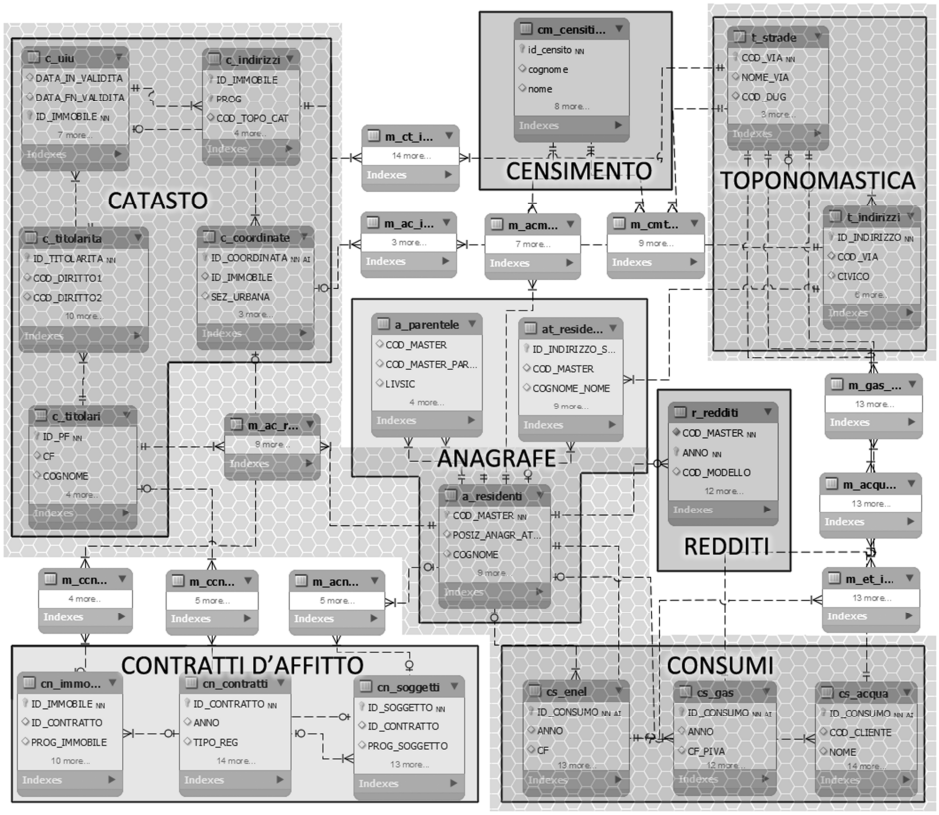

Figura 14 - Schema dati del database riconciliato

Nello schema, le aree col pattern esagonale di sottofondo evidenziano ciò che è rimasto intatto dalla situazione iniziale (vedi paragrafo 3.1.5). Inoltre, nel model-

lo logico sono indicate le sole tabelle principali del database. Per motivi di spazio, sono state omesse le tabelle che contengono solamente le decodifiche degli acronimi e delle sigle utilizzate in altre tabelle; lo stesso discorso è valido per le varie tabelle temporanee e/o di supporto create per le operazioni di match.

4 Il progetto DyNamiTE: il front-end

La seconda parte del progetto si concentra sullo sviluppo del front-end del sistema, ossia delle funzionalità che possono essere realizzate sfruttando il database riconciliato costruito. In primo luogo sono stati affrontati i pattern di evasione, i quali costituiscono l'obiettivo principale del progetto. In seguito si è provveduto alla realizzazione di un'interfaccia, per rendere fruibile il database riconciliato da parte degli utenti. Un'ultima finestra è stata invece dedicata allo sviluppo di una funzionalità alternativa alla lotta all'evasione fiscale, per dare dimostrazione del potenziale del database riconciliato.

4.1 I pattern di evasione

La lotta digitale all'evasione fiscale prevede di sfruttare l'integrazione di banche dati diverse per analizzare un complesso insieme di informazioni e individuare le sospette circostanze fraudolente. Tale attività di analisi prende forma con i cosiddetti *pattern di evasione*, ossia quei percorsi di navigazione dello schema riconciliato che, selezionando i dati in base a determinate caratteristiche, permettono di rilevare le suddette circostanze. La costruzione di un pattern di evasione comincia dalla formulazione di un'ipotesi: in sostanza, bisogna inquadrare una specifica situazione sospetta da ricercare e definire quali siano i parametri (ossia le caratteristiche dei dati) attraverso cui tale situazione possa essere rilevata. Il passo successivo consiste nel testare l'ipotesi formulata, verificandone la validità e apportando eventuali modifiche ai parametri per migliorarne l'affidabilità. Una volta conclusa la ricerca del pattern, i risultati vengono sottoposti all'attenzione dei membri

dell'Ufficio Tributi del Comune, con lo scopo di avere una conferma sull'effettivo sospetto dei casi rilevati o per ricevere indicazioni su come migliorare o correggere i parametri di ricerca. Siccome la maggior parte delle informazioni è disponibile per l'anno 2010, tutti i pattern di evasione concentrano le loro ricerche su questa annata.

4.1.1 Consumi fuori soglia

Il primo pattern su cui si è deciso di concentrarsi riguarda l'individuazione dei consumi fuori soglia. L'obiettivo consiste sostanzialmente nel rilevare le utenze che presentano un consumo sproporzionato rispetto al numero di occupanti; in tali casi, il consumo elevato potrebbe essere dovuto alla presenza di un maggior numero di occupanti e, di conseguenza, ci si potrebbe trovare di fronte a degli affitti in nero. Un'altra ipotesi formulabile è invece che in una UIU di tipo residenziale si nasconda un'attività commerciale. Sebbene il pattern sia concettualmente semplice, a livello pratico è necessaria una lunga serie di step intermedi.

1. **Calcolo dell'occupazione media** in un indirizzo di residenza nel corso del 2010. A tal proposito è stata creata una tabella temporanea, *at_residenti_indirizzi_2010*, in cui sono state riportate le residenze valide nel corso del 2010. A tali residenze sono state modificate le date di validità, per farle rientrare entro i limiti del 2010 (e facilitare così il calcolo successivo). Dopodiché è stata creata un'altra tabella, *t_indirizzi_2010_occupazione*, in cui ad ogni indirizzo è stata associata l'occupazione media nel corso del 2010.
2. **Associazione delle utenze agli indirizzi**. Data l'assenza della numerazione civica interna negli indirizzi delle utenze, le associazioni vengono stabilite secondo due criteri:
    - Se l'utenza si trova in un indirizzo in cui non ci sono civici interni, l'associazione utenza-indirizzo è immediata.
    - In caso contrario (se a tale indirizzo sono invece presenti più civici interni), l'associazione avviene soltanto se l'intestatario dell'utenza ha risieduto nel 2010 in un indirizzo che coincide con quello dell'utenza stessa. E' possibile che una persona sia intestataria di più utenze in una stessa palazzina; in questi casi,

all'indirizzo di residenza di tale persona viene associata l'utenza con consumo più alto.

Al termine dell'operazione si hanno tre tabelle temporanee (*_cs_enel_uiu*, *_cs_acqua_uiu* e *_cs_gas_uiu*) in cui sono memorizzati l'ID dell'indirizzo, l'ID dell'utenza associata e il consumo dell'utenza.

3. **Calcolo dei percentili dei consumi** in base al numero di occupanti. Le tre tabelle create al punto precedente sono state esportate in tre file Excel, prendendo - per ogni associazione utenza-indirizzo - il consumo dell'utenza e l'occupazione media dell'indirizzo (calcolata al punto 1). Per ogni tipo di utenza, i consumi sono stati suddivisi in base al numero di occupanti e ordinati per valori crescenti del consumo; per ogni "numero di occupanti" sono stati quindi selezionati i consumi corrispondenti a determinati percentili.

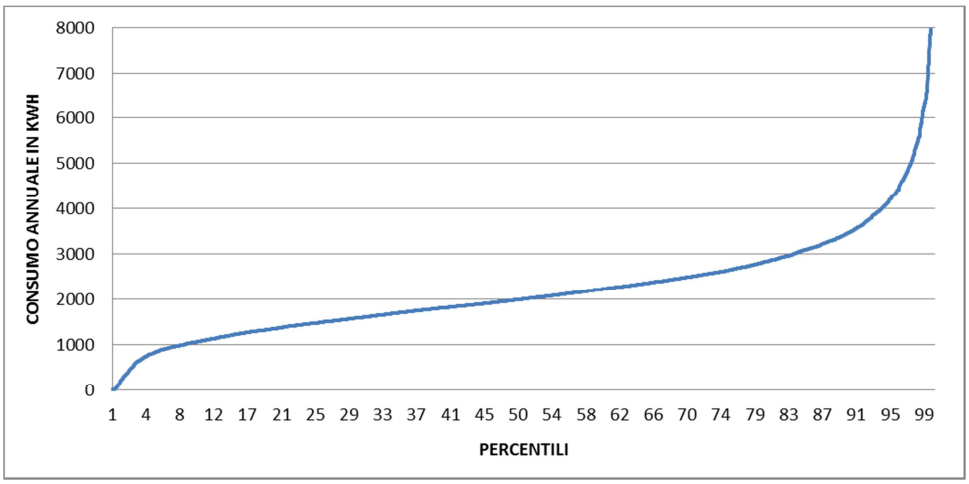

Figura 15 - Percentili delle utenze elettriche associate ad un'occupazione media di due persone

Il grafico in Figura 15 mostra i consumi (associati ad un'occupazione media di 2 persona) ordinati per valori crescenti e la corrispondente scala dei percentili: per "n° percentile" si intende il consumo rilevato nella posizione corrispondente all'n% dell'ordinamento. Per ogni valore di "occupazione media" nelle utente elettriche e di gas sono stati rilevati i percentili più significativi (1, 2, 3, 4, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 95, 96, 97, 98, 99), dopodiché i risultati sono stati inseriti nel database riconciliato nelle tabelle *cs_soglie_enel* e *cs_soglie_gas*.

Per quanto riguarda le utenze dei consumi d'acqua, il calcolo dei percentili è leggermente diverso: i consumi sono solamente espressi in fasce di consumo (per un totale di 6 fasce), perciò - ad esempio - la fascia di consumo al 99-esimo percentile può coincidere con la fascia di consumo all'80-esimo. Per questo motivo si è deciso di scegliere i soli percentili corrispondenti agli "scatti" di fascia.

Un'ultima osservazione va fatta sulla distribuzione dei percentili: si è infatti notato il fenomeno per cui, al crescere del numero di occupanti, i valori dei percentili non sempre crescono, ma talvolta risultano diminuire. Tale fenomeno, che si verifica principalmente in corrispondenza di un numero elevato di occupanti (da 7 in su), può essere spiegato dalla scarsa numerosità delle utenze associate per le suddette fasce di occupanti.

4. **Calcolo del numero di occupanti stimato** in base ai consumi. Data la distribuzione dei consumi in base all'occupazione media, è possibile fare l'operazione inversa per stimare l'occupazione media in base al valore del consumo. In particolare, l'occupazione stimata viene determinata in base alla minor distanza dai 50-esimi percentili dei consumi; in altre parole, presi i 50-esimi percentili associati ai diversi numeri di occupanti, il risultato corrisponde al numero di occupanti del 50-esimo percentile a cui il consumo è più vicino. Data la distribuzione non lineare dei percentili citata al punto precedente, sono stati considerati i soli 50-esimi percentili che presentano un andamento crescente (ossia da 1 a 6 occupanti).

   Per quanto riguarda le utenze dei consumi d'acqua, la stima del numero di occupanti è avvenuta in maniera diversa (per lo stesso motivo citato al punto precedente). Date tutte le utenze per una specifica fascia di consumo, la stima del numero di occupanti è stata scelta come la moda del numero di occupanti associati a tali utenze.

5. **Associazione alle utenze dei dati calcolati** nei punti 3 e 4. Sono state create delle nuove tabelle temporanee (*_cs_enel_perc, _cs_acqua_perc* e *_cs_gas_perc*) in cui - ad ogni coppia indirizzo-

utenza - sono stati associati il percentile del consumo e il numero stimato di occupanti.

6. **Creazione della tabella finale**, *cs_consumi_indirizzi*, in cui vengono riassunti tutti i dati calcolati per ogni indirizzo. Va precisato che gli indirizzi elencati in questa tabella comprendono solo quelli a cui è stata associata almeno un'utenza. In particolare, la tabella contiene le seguenti informazioni:
    - ID dell'indirizzo
    - Numero medio di occupanti
    - Per ogni tipo di utenza eventualmente associata:
        - ID dell'utenza
        - Percentile del consumo
        - Numero stimato di occupanti
    - Media del numero stimato di occupanti per ogni utenza.

Al termine di questa lunga procedura, il pattern è praticamente immediato: è infatti sufficiente considerare gli indirizzi che presentano consumi superiori ad un determinato percentile di soglia. Sebbene il pattern presenti delle situazioni interessanti, nessuna di queste è stata effettivamente approfondita: basandosi sulla sola presenzi di consumi elevati, l'assunzione dell'affitto in nero può risultare abbastanza flebile. Ciò non significa che le informazioni calcolate siano inutili: i consumi troppo elevati sono sicuramente sospetti a priori, ma sono necessari maggiori parametri per approfondire e valutare meglio le varie situazioni. Il prossimo pattern concretizza specificamente questa necessità.

4.1.2 Falsi separati

Il pattern dei falsi separati nasce come evoluzione del pattern dei consumi fuori soglia: si è infatti voluto cercare un contesto in cui la rilevazione di un consumo sproporzionato potesse indicare con maggior certezza il verificarsi di un affitto in nero. Quello dei falsi separati costituisce uno specifico scenario di evasione, in cui due coniugi convivono, possiedono due immobili e mascherano un affitto in nero sulla seconda casa dichiarando due residenze separate (una nella casa in cui abitano veramente e l'altra nella seconda casa). Per capire se si tratti veramente della situazione descritta o se si tratti invece di una normale separazione è necessario

confrontare i consumi sulle due abitazioni; in particolare, sono due le casistiche che possono destare sospetti:

Tabella 8 - Descrizione delle casistiche considerate nel pattern dei falsi separati

Casistica	Descrizione
"Alto-basso"	In una delle abitazioni si registra un consumo elevato rispetto al numero di occupanti, mentre nell'altra abitazione si registra un consumo molto basso. L'ipotesi è quindi che la seconda casa sia lasciata sfitta; in questo modo, la dichiarazione delle residenze separate permetterebbe di evadere il pagamento della tassa ICI sulla seconda casa.
"Alto-alto"	In entrambe le abitazioni si registra un consumo elevato rispetto al numero di occupanti. In questo caso, l'ipotesi è che sulla seconda casa sia in atto un affitto in nero - al quale si aggiungerebbe, come nel caso precedente, l'evasione dell'ICI.

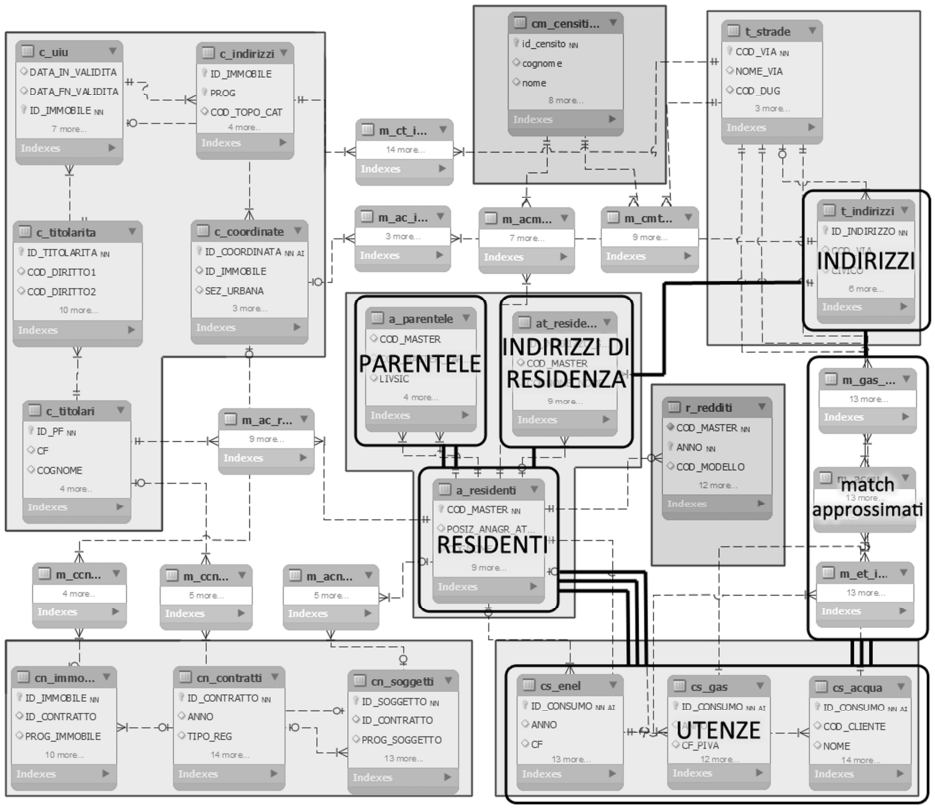

Figura 16 - Il percorso delle relazioni coinvolte nel pattern dei falsi separati

La logica alla base di questo percorso di navigazione può essere scomposta nei punti seguenti:

1. **Calcolo dei consumi medi negli indirizzi di residenza** in base al numero di occupanti. Tale operazione è stata eseguita nel corso del pattern precedente e si è conclusa con l'assegnazione di un percentile ad ogni utenza (rispetto al numero di occupanti).
2. **Individuazione delle coppie separate**, ossia persone sposate entro il 2010 e che, nel corso del 2010, risultano residenti in indirizzi diversi (ma sempre nel Comune di Cesena). A tal proposito è stata creata una tabella, *pt_spac_2010*, in cui sono stati memorizzati gli ID di entrambi i coniugi e gli ID dei rispettivi indirizzi di residenza.
3. **Selezione e classificazione delle coppie** in base ai parametri del pattern, ossia la discordanza dei consumi rispetto al numero di occupanti. Innanzitutto si definiscono i percentili che costituiscono le "soglie di sospetto", cioè oltre i quali i consumi vengono definiti "troppo alti" o "troppo bassi"; di base sono stati scelti il 15° percentile come soglia per i consumi bassi e l'85° per quelli alti. A questo punto, per ogni coppia individuata al punto precedente vengono presi i percentili delle utenze agganciate sui rispettivi indirizzi di residenza e si calcola il "punteggio di sospetto". Il calcolo del punteggio (di seguito descritto tramite pseudocodice) si differenzia in base al tipo di pattern ricercato.

```
/* Pattern alto-alto */
foreach (coppia) {
    punteggio = 0; numeroUtenze = 0;
    foreach (utenza in utenzeDellaCoppia) {
            numeroUtenze++;
            if (percentileDellUtenza >= sogliaAlta)
                    punteggio += percentileDellUtenza;
    }
    punteggio /= numeroUtenze;
}

/* Pattern alto-basso */
foreach (coppia) {
    punteggio = 0; numeroUtenze = 0;
    foreach (utenza in utenzeDelMarito) {
            numeroUtenze++;
            if (percentileDellUtenza >= sogliaAlta)
                    punteggio += percentileDellUtenza;
    }
    foreach (utenza in utenzeDellaMoglie) {
            numeroUtenze++;
            if (percentileDellUtenza <= sogliaBassa)
                    punteggio += 100 - percentileDellUtenza;
    }
    punteggio /= numeroUtenze;
}
/* Il punteggio risulta variabile da 0 (nessun consumo fuori soglia) a 99 (tutti i consumi
"sballati"). Le coppie vengono quindi ordinate per punteggio decrescente. */
```

Figura 17 - Pseudocodice per il calcolo del punteggio di sospetto dei falsi separati

Va sottolineato che, ai fini del pattern, è importante che i due coniugi (o qualche loro parente) siano titolari dei due rispettivi immobili: è infatti ovvio che, nel caso uno dei due coniugi sia in affitto, l'ipotesi di sospetto cade automaticamente. Tuttavia, a causa della non elevata numerosità dei casi rilevati e delle difficoltà riscontrate negli agganci tra gli indirizzi di residenza e gli immobili, questo tipo di controllo è stato demandato alla fase di revisione manuale dei singoli casi: il rischio di perdere dei casi interessanti sarebbe infatti maggiore del rischio di includere qualche falso positivo.

A partire da una popolazione di 44860 persone attualmente vive e sposate nel 2010, quelle che hanno risieduto in abitazioni diverse nel corso del 2010 corri-

spondono a 1618. Per ognuna di queste sono state cercate le titolarità catastali e i consumi sull'indirizzo di residenza.

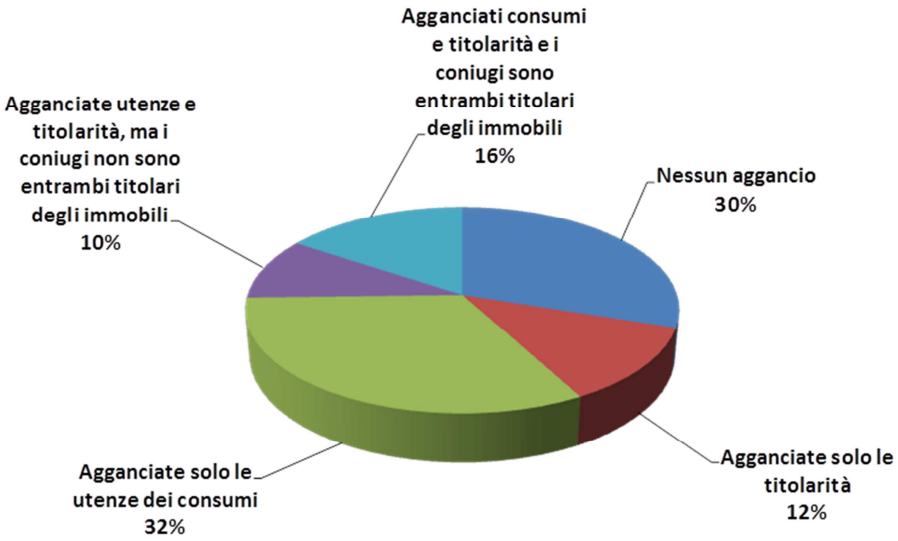

Figura 18 - Statistiche sul campione dei potenziali falsi separati

Da questa distribuzione dei dati emerge che solo il 16% dei candidati hanno le caratteristiche necessarie per applicare appieno il pattern. Se il 10% si può escludere per non aderire ai parametri richiesti, il restante 74% non può essere completamente approfondito a causa degli agganci mancanti o soltanto parziali. E' in queste percentuali che si riflettono i problemi principalmente dovuti all'assenza dei civici interni nelle titolarità catastali e nelle utenze; una qualità migliore dei dati avrebbe sicuramente permesso di attingere ad un bacino maggiore di potenziali evasori. Tuttavia, nonostante i problemi citati, i dati a disposizione sono stati sufficienti per testare la validità del pattern. Per ogni tipo di pattern (alto-alto e alto-basso) sono stati selezionati e manualmente controllati i primi 10 casi: alcuni - come previsto - si sono rivelati falsi positivi, o perché un coniuge risultava in affitto, o perché i conviventi sui due indirizzi di residenza rendevano palese la separazione; su molti sono stati invece confermati i sospetti. Il caso di seguito esposto costituisce un esempio dei falsi separati che il pattern ha permesso di rilevare. Per motivi di privacy, i nomi delle persone e gli indirizzi di residenza non vengono riportati.

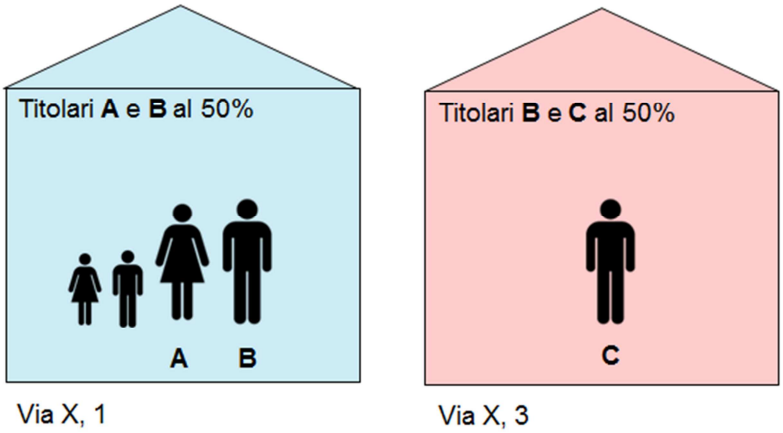

Figura 19 - Falsi separati: un esempio (situazione iniziale)

Nella prima casa si trova una famiglia di quattro persone, composta da moglie (*A*), marito (*B*) e due figli; i due coniugi risultano inoltre co-proprietari dell'immobile. Nella casa adiacente si trova invece un single di mezza età (*C*), senza legami di parentela coi vicini; la proprietà dell'immobile è condivisa dal residente e dal marito della famiglia accanto. Questa situazione resta invariata fino al 2006, quando *C* cambia residenza e lascia la sua titolarità dell'immobile a *B*. In corrispondenza di questo avvenimento, *B* sposta la residenza nella casa adiacente (lasciata vuota da *C*) e cede alla moglie la titolarità dell'immobile in cui risiede il resto della famiglia. La situazione resta poi immutata fino alla data odierna; va sottolineato che, a più di 6 anni dalla separazione delle residenze, i due coniugi non hanno mai divorziato.

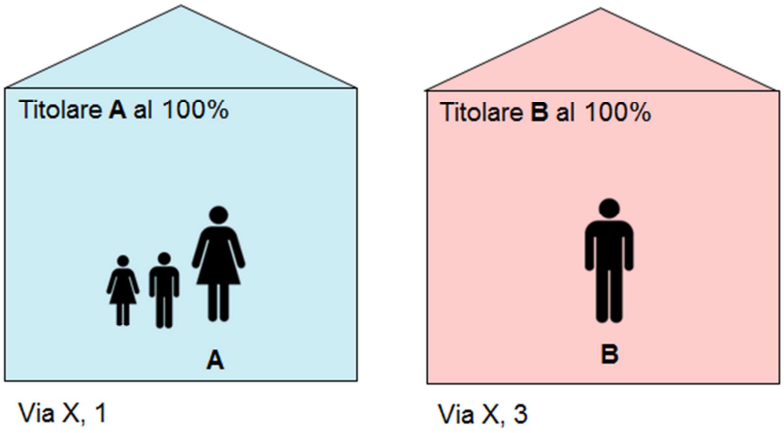

Figura 20 - Falsi separati: un esempio (situazione attuale)

L'elemento del pattern che ha fa scattare il sospetto su questo caso è il consumo rilevato nelle abitazioni nel 2010: nella prima abitazione viene rilevato un

consumo di elettricità (intestato al marito) compatibile con 3-4 occupanti, mentre nella seconda casa vengono rilevati consumi di gas e di acqua (intestati sempre al marito) consoni ad un'occupazione di 6 persone e assolutamente sproporzionati per un solo occupante. In conclusione, è ragionevole supporre che il marito abbia continuato ad abitare col resto della famiglia e che la seconda casa sia stata effettivamente affittata ad un gruppo di persone.

I casi sospetti confermati a seguito di un approfondimento manuale (10 in tutto) sono stati sottoposti alla validazione dell'Ufficio Tributi: il personale di questo ufficio svolge abitualmente controlli di tipo tributario e ha l'esperienza per poter valutare la sospettosità di un caso. Il risultato della loro analisi è stato sicuramente buono: la metà dei casi sottoposti sono stati identificati come "fortemente sospetti" e uno di questi è stato immediatamente segnalato all'Agenzia delle Entrate. Il carico di lavoro dell'Ufficio Tributi non ha permesso agli addetti di approfondire tutti i casi, ma il feedback ricevuto è stato positivo e ha confermato la validità del pattern implementato.

4.1.3 Analisi su patrimonio economico e immobiliare

L'idea per questo pattern nasce dalla volontà di sfruttare le dichiarazioni dei redditi per scoprire altri evasori. Se fossero stati disponibili i dati completi (e fossero stati quindi conosciuti l'utilizzo dichiarato e l'eventuale locazione di ogni immobile posseduto), le ricerche possibili sarebbero state molte. Potendo invece usufruire delle sole dichiarazioni sintetiche, si è reso necessario lo studio di casistiche elaborate. In particolare, si è deciso di basare le ricerche sul confronto del patrimonio economico di una persona (derivante dalle dichiarazioni dei redditi) con il suo patrimonio immobiliare (calcolato dai dati catastali; per i dettagli, consultare il paragrafo 3.3.4). Il fulcro di questo tipo di ricerche è costituito dai seguenti dati:

- **Il rapporto tra il reddito totale e la rendita catastale totale.** La supposizione è che, in generale, ad una rendita catastale elevata debba corrispondere un reddito elevato.
- **Il rapporto tra il reddito da fabbricati e la rendita catastale totale.** Il reddito di ogni immobile è costituito dal reddito dovuto al possesso (pari alla rendita catastale aumentata del 5%) sommato al reddito do-

vuto alle eventuali locazioni. Per questo motivo, si suppone che - in assenza di contratti d'affitto - il reddito da fabbricati debba essere leggermente maggiore alla rendita catastale (con un rapporto prossimo di 1,05).

Va precisato che la rendita catastale totale è stata calcolata in due modi diversi: sia pesando la rendita di ogni immobile sulle quote e sulle durate del possesso, sia considerandone il possesso completo per tutto l'anno. Il primo calcolo dovrebbe essere ovviamente più affidabile, ma l'assenza e l'incongruenza di molte quote di possesso può produrre un risultato poco affidabile. Per questo motivo si è deciso di mantenere entrambi i calcoli; di riflesso, anche i rapporti sopra menzionati sono stati calcolati con entrambe le rendite catastali ottenute. Basandosi sui patrimoni economici e immobiliari delle persone si è deciso di ricercare tre diverse situazioni:

1. **Incongruenza tra il reddito e la rendita catastale**. Si vogliono individuare quei casi in cui ad una rendita catastale complessivamente elevata non corrisponde un reddito totale elevato; anche se gli immobili sono stati ereditati, si ritiene necessario un buon reddito per poterli mantenere. Si cercano quindi le persone per cui il rapporto tra reddito totale e rendita catastale è minore di 1.
2. **Immobili non dichiarati**. Si vogliono individuare quei casi in cui il rapporto tra il reddito da fabbricati e la rendita catastale è minore di 1, ad indicare la possibilità che il possesso di alcuni immobili non sia stato dichiarato.
3. **Affitti non dichiarati**. In questo caso, l'assunzione è un po' più elaborata: sapendo che il rapporto tra il reddito da fabbricati e la rendita catastale totale dovrebbe corrispondere a 1,05 in assenza di locazioni, si vogliono cercare quei casi in cui il suddetto rapporto non sembra indicare affitti (cioè il rapporto rientra nell'intorno di 1,05), ma dai dati sugli immobili si può pensare che almeno uno di questi sia stato dato in affitto. L'ipotesi della presenza di un affitto si basa sul numero di fabbricati residenziali abitati (campo già calcolato e presente in *r_redditi*): se i fabbricati residenziali abitati sono più di uno (si esclude chiaramen-

te quello in cui risiede il titolare), allora è probabile che qualcuno di questi sia stato affittato. Riassumendo, i parametri richiesti sono i seguenti:

- Rapporto tra il reddito da fabbricati e la rendita catastale totale compreso tra 1 e 1.5 (per garantire un minimo di tolleranza).
- Due o più fabbricati residenziali posseduti e abitati.

Inoltre, a causa della delicatezza dell'assunzione, è preferibile aggiungere i seguenti parametri (anch'essi corrispondenti a campi già calcolati e presenti in *r_redditi*) per avere una maggior affidabilità sui risultati ottenuti:

- Tutti i fabbricati devono essere stati posseduti interamente e per tutto l'anno.
- Su nessuno dei fabbricati posseduti devono esistere incongruenze sulle quote di possesso.

Per ciascuna delle tre situazioni descritte, la ricerca dei casi sospetti avviene semplicemente filtrando i record della tabella *r_redditi* sui parametri indicati. Diversamente dal pattern dei falsi separati, in questo caso non è previsto il calcolo di un punteggio sospetto, ma ci si limita ad ordinare i risultati sui campi interessanti (ossia quelli su cui si basa la selezione). Dalle ricerche su tutti e tre gli scenari sono stati estratti un totale di 25 casi potenzialmente interessanti; uno di questi viene di seguito riportato come esempio:

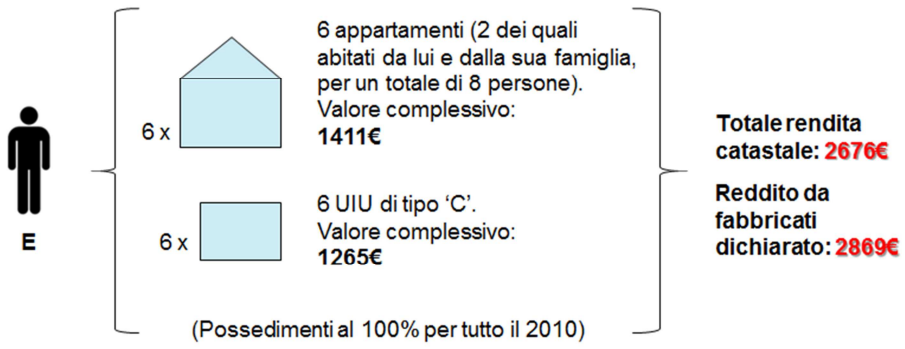

Figura 21 - Analisi su patrimonio economico ed immobiliare: un esempio

Come si può notare dalla Figura 21, il soggetto riportato risulta proprietario al 100% di 6 appartamenti e di 6 UIU di tipo C (garage, cantine, ecc.) e il reddito da fabbricati dichiarato è compatibile con il reddito dovuto per la sola proprietà degli

immobili. Tuttavia, è ragionevole supporre che una persona titolare di 5 seconde case non le lasci tutte sfitte, perciò si è deciso di effettuare degli approfondimenti manuali sugli indirizzi degli immobili. Dall'indagine si è rilevato che:

- Una delle seconde case è abitata da parenti del soggetto; si può quindi escludere l'affitto dell'immobile.
- In un'altra delle seconde case risulta residente una donna straniera, per un periodo di circa 4 mesi nel corso del 2010.
- Un'altra ancora delle seconde case risulta associata ad utenze di acqua e gas i cui consumi sono compatibili con un'occupazione di almeno 6 persone.
- Per le altre due seconde case non sono stati rilevati residenti o utenze, ma i mancati agganci potrebbero essere dovuti al problema del civico interno degli appartamenti piuttosto che all'effettiva assenza di residenti e di utenze.

Da questa indagine si può concludere che, con buona probabilità, il soggetto *E* abbia effettivamente affittato almeno due appartamenti; il fatto che tali affitti siano in nero sarebbe dimostrato dal reddito da fabbricati, il cui importo sarebbe dovuto essere superiore al reddito dovuto per il possesso degli immobili.

I 25 casi estratti sono stati sottoposti al personale dell'Ufficio Tributi, per ricevere un responso sulla validità del pattern. Sebbene non tutti siano stati esaminati, le analisi da loro effettuate sono state sufficienti per generare un feedback correttivo: il pattern si basa sull'assunzione che il reddito da fabbricati e la rendita catastale debbano avere un rapporto di 1,05, ma questo non è sempre corretto. Ciò è dovuto al fatto che alcune categorie di fabbricati non costituiscono reddito, perciò le rendite catastali di questi immobili non devono essere dichiarate; di queste categorie fanno parte i capannoni rurali, le cui rendite catastali sono spesso molto elevate e che, senza i dovuti controlli, possono generare dei falsi positivi. A causa dei tempi stretti, la correzione al meccanismo del pattern non è stata apportata, perciò la stessa viene demandata a sviluppi futuri. Tuttavia, la modifica prevista non dovrebbe essere impegnativa: sarebbe infatti sufficiente creare un nuovo campo in cui calcolare la rendita catastale per i soli immobili che producono reddito. Di conseguenza, il reddito da fabbricati andrebbe confrontato con questo nuo-

vo campo anziché con la pura rendita catastale totale; il resto del pattern dovrebbe invece restare invariato.

4.2 L'interfaccia

Uno degli obiettivi del progetto è quello di rendere fruibili il database riconciliato e i pattern di evasione al personale del Comune di Cesena (in particolare ai membri dell'Ufficio Tributi). Si è pertanto deciso realizzare un'applicazione web, da mettere a disposizione sulla intranet del Sistema Informativo e che offra un'interfaccia semplice, intuitiva e funzionale. Lo sviluppo della suddetta interfaccia è stato agevolato dalla possibilità di usufruire di un framework interno, sviluppato dal personale del Sistema Informativo appositamente per la realizzazione di applicazioni web da collocare sulla intranet. Tale framework permette di gestire con assoluta facilità la problematica dei permessi di accesso per gli utenti e implementa automaticamente le funzionalità di login; inoltre, esso fornisce una vasta serie di vantaggi quali la standardizzazione del layout di base, delle chiamate AJAX e della rappresentazione dei risultati di un'interrogazione SQL. La somma di tutti questi fattori ha consentito la realizzazione dell'applicazione in tempi molto rapidi.

4.2.1 La metafora adottata

Il nucleo dell'applicazione è costituito dalla possibilità di navigare il database riconciliato attraverso un'interfaccia grafica; si vuole cioè dare agli utenti un modo semplice ed intuitivo che permetta di consultare le informazioni e di "spostarsi" a piacimento all'interno del database sfruttando l'integrazione dei dati. A tal proposito sono stati individuati nel database riconciliato i concetti fondamentali su cui basare la navigazione.

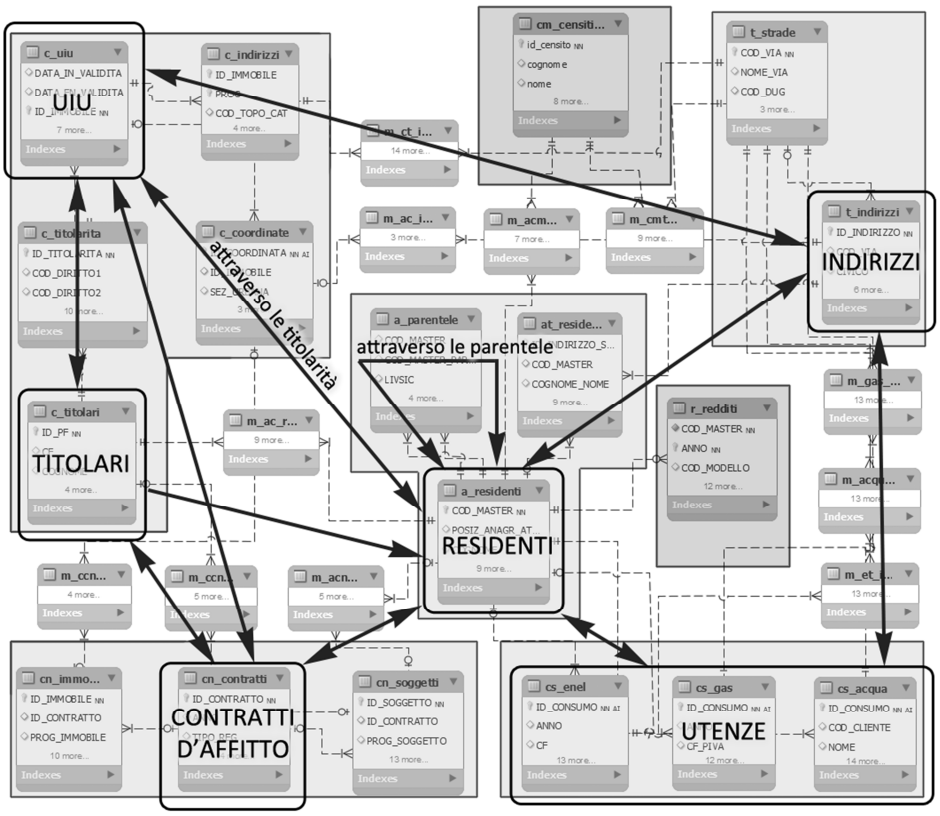

Figura 22 - Meccanismo di navigazione dei concetti dello schema

Nello schema, ognuno degli elementi cerchiati costituisce un concetto consultabile, mentre le frecce indicano le direzioni degli spostamenti da un concetto all'altro. La consultazione di un'istanza di questi concetti deve consentire di visualizzare tutte le informazioni disponibili sulla data istanza e di fornire il collegamento alle istanze agganciate nei concetti adiacenti. Per raggiungere tale obiettivo si è resa necessaria la progettazione di una metafora in grado di fornire un'interfaccia intuitiva, facile da comprendere e da utilizzare. Per la descrizione della metafora adottata viene di seguito riportato uno screen dell'interfaccia; per motivi di privacy, i dati sono stati ovviamente offuscati.

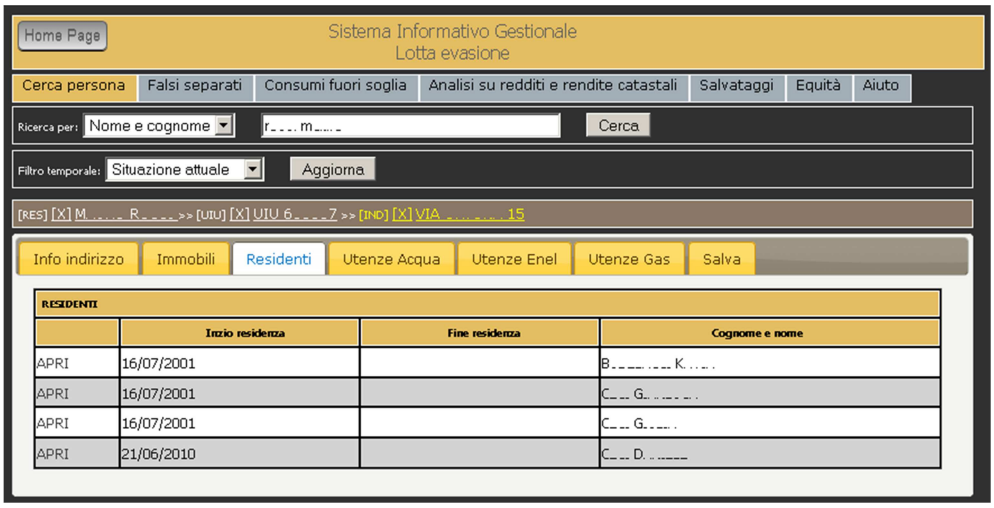

Figura 23 - Screen dell'interfaccia: visualizzazione della scheda di un indirizzo

Le lettere riportate sul lato sinistro dell'immagine sono state aggiunte per facilitare la seguente descrizione dei componenti dell'interfaccia.

Tabella 9 - Descrizione dei componenti dell'interfaccia

Comp.	Descrizione
A	Header dell'applicazione.
B	**Barra del menù**: consente di scegliere una delle funzionalità implementate.
C	**Barra di ricerca**: consente all'utente di avviare una nuova ricerca, i cui parametri sono relativi alla funzionalità selezionata.
D	**Barra del filtro temporale:** consente all'utente di scegliere la finestra temporale dei dati da visualizzare; le opzioni sono: • Situazione attuale. • Storico completo. • Situazione valida in un determinato anno. • Situazione valida in un arco compreso tra due date specificate. Tali opzioni sono modificabili in corso di navigazione.

E	**Barra del percorso:** consente di visualizzare il percorso di navigazione finora seguito dall'utente. Nell'esempio in Figura 23, si può vedere come l'utente sia partito dai dati di un residente, sia passato ai dati di un immobile da esso posseduto e sia poi giunto all'indirizzo in cui si trova l'immobile. La visualizzazione dei dati precedentemente consultati è fondamentale per tener traccia del percorso seguito, specialmente se si considera il fatto che la navigazione è potenzialmente infinita. Inoltre, gli elementi riportati in questa barra possono essere selezionati per consultare i dettagli della relativa istanza - il tutto senza perdere traccia del percorso. Qualora l'utente decida di "diramare" la navigazione (ossia decida di consultare una nuova istanza a partire da una delle istanze precedentemente consultate), il percorso di navigazione viene sempre mantenuto intatto e la nuova istanza viene aggiunta in fondo alla barra del percorso. Infine, ogni elemento sulla barra del percorso è dotato di una "[X]" che, se cliccata, consente l'eliminazione del suddetto elemento dalla barra (senza però modificare il resto del percorso).
F-G	**Scheda dell'istanza:** in quest'area vengono mostrate tutte le informazioni disponibili sull'istanza attualmente selezionata. All'inizio della navigazione (che comincia solitamente con una ricerca), quest'area viene dedicata alla presentazione (in forma tabellare) delle istanze da cui è possibile avviare la navigazione.
F	**Tab delle informazioni disponibili:** le informazioni collegate all'istanza vengono raggruppate in base alla provenienza. Il primo tab fornisce solitamente le informazioni specifiche dell'istanza (ad esempio, i dati anagrafici dal residente), mentre gli altri tab forniscono le informazioni sui concetti collegati (ad esempio, gli immobili posseduti dal residente, le utenze intestate, l'elenco dei parenti, ecc.).
G	**Area delle informazioni:** i dati relativi al tab selezionato vengono solitamente mostrati in formato tabellare. Qualora i dati mostrati siano relativi ad un'istanza navigabile (ad esempio un immobile, un'utenza, ecc.), la prima colonna contiene i link "APRI" che consentono di consultare l'istanza selezionata; questa operazione - implementata con una chiamata AJAX - causa il ricaricamento dell'area dell'istanza (F-G) con le nuove informazioni e l'aggiornamento della barra del percorso (E).

4.2.2 Le funzionalità di navigazione

La metafora implementata per la navigazione dei dati ha consentito lo sviluppo di diverse funzionalità da mettere a disposizione degli utenti. Innanzitutto si è deciso di offrire la **libera navigazione** dello schema riconciliato; questa funzionalità è stata fortemente voluta dai membri dell'Ufficio Tributi, interessati a sfruttare le potenzialità del database riconciliato per le loro attività quotidiane di controllo. Come punto di partenza della navigazione sono stati scelti i residenti, che - come si può vedere in Figura 22 - costituiscono il concetto più importante, nonché collegato a tutti gli estremi del database. La suddetta funzionalità corrisponde alla voce "Cerca persona" nella barra del menù; selezionando questa opzione si accede ad una semplice form, tramite cui è possibile cercare una persona in base al codice fiscale o al nome e cognome.

Figura 24 - Screen dell'interfaccia: ricerca iniziale per avviare la libera navigazione

Come si può vedere in Figura 24, l'elenco dei residenti che rispondono ai parametri inseriti viene mostrato in forma tabellare, completo di tutti i rispettivi dati anagrafici (in modo da poter scegliere la persona giusta nei casi di omonimia). La prima colonna della tabella contiene i link "APRI", i quali consentono di avviare la navigazione del database a partire dal residente sulla riga selezionata. A questo punto, l'utente può consultare tutte le informazioni sul residente selezionato ed è libero di muoversi nel database riconciliato sfruttando la metafora implementata.

Lo sviluppo di questa funzionalità è servito anche come base per la realizzazione delle funzionalità successive, ossia quelle dedicate all'esecuzione dei **pattern di evasione**. Per ciascuna di esse viene infatti replicato lo stesso meccanismo: attraverso una form vengono inseriti i parametri del pattern scelto e i risultati vengono mostrati nella stessa forma tabellare; dopodiché, selezionando uno dei link "APRI"

si accede alla navigazione dei dati a partire dall'istanza selezionata. La lista dei parametri e l'elenco dei risultati risultano chiaramente diversi a seconda del pattern selezionato. Il primo pattern implementato è quello dei falsi separati, selezionabile dall'omonima voce nella barra del menù. I parametri richiesti per avviare la ricerca sono i seguenti:

Tabella 10 - Parametri richiesti per il pattern dei falsi separati

Parametro	Descrizione
Anno	Al momento è disponibile solo il 2010
Tipo di pattern	"Alto-alto" per cercare i coniugi a cui sono collegati consumi elevati nelle rispettive residenze. "Alto-basso" per cercare i coniugi a cui sono collegati un consumo alto in una residenza e un consumo basso nell'altra.
Percentile basso	Valore di soglia per identificare un consumo come "basso". Deve essere compreso tra 1 e 50; di default è 15.
Percentile alto	Valore di soglia per identificare un consumo come "alto". Deve essere compreso tra 50 e 99; di default è 85.

La procedura di elaborazione del pattern e di individuazione dei sospetti è stata implementata in un file PHP, al quale vengono inviati i parametri inseriti nella form. Al termine dell'elaborazione, i risultati vengono mostrati in una forma tabellare che rispecchia il seguente formato:

Tabella 11 - Formato dei risultati del pattern dei falsi separati

Campo	Descrizione
'APRI'	Link per avviare la navigazione.
Coniuge 1	Nome e cognome del primo coniuge.
Percentile Enel 1	Eventuale percentile del consumo elettrico rilevato.
Percentile Acqua 1	Eventuale percentile del consumo di acqua rilevato.
Percentile Gas 1	Eventuale percentile del consumo di gas rilevato.
Coniuge 2	Nome e cognome del secondo coniuge.
Percentile Enel 2	Eventuale percentile del consumo di elettrico rilevato.
Percentile Acqua 2	Eventuale percentile del consumo di acqua rilevato.

Percentile Gas 2	Eventuale percentile del consumo di gas rilevato.
Punteggio	Valore del punteggio di sospetto associato alla coppia.

I risultati mostrati sono chiaramente ordinati per valori decrescenti di punteggio. Selezionando il link "APRI", l'utente viene rimandato ad una scheda riassuntiva della coppia, in cui vengono fornite le informazioni essenziali dei due coniugi (dati anagrafici e indirizzi di residenza) e i link per avviare la navigazione su uno dei due soggetti. Per quanto riguarda il pattern dei consumi fuori soglia (anch'esso selezionabile dall'omonima voce nella barra del menù), i parametri richiesti sono i seguenti:

Tabella 12 - Parametri richiesti per il pattern dei consumi fuori soglia

Parametro	Descrizione
Anno	Al momento è disponibile solo il 2010
Consumi da considerare	Tre checkbox ("Enel", "Acqua" e "Gas") permettono di scegliere i consumi su cui eseguire le ricerche.
Tipo di ricerca	Selezionabile tra "Consumo troppo alto" e "Consumo troppo basso".
Percentile di soglia	Valore di soglia dei percentili; il fatto che si intenda "soglia alta" o "soglia bassa" dipende dal tipo di ricerca. Deve essere compreso tra 1 e 99.

L'elaborazione del pattern prevede la semplice estrazione degli indirizzi di residenza a cui sono associati consumi fuori dalla soglia specificata. I risultati vengono quindi mostrati nella solita forma tabellare, in cui compaiono (oltre al link "APRI") gli estremi dell'indirizzo di residenza, i percentili dei consumi eventualmente associati e il numero medio di occupanti durante l'anno. Infine, i risultati vengono ordinati in maniera decrescente rispetto alla somma dei percentili dei consumi associati. Il terzo ed ultimo pattern è selezionabile dall'etichetta "Analisi su redditi e rendita catastale". Per questo pattern, l'elenco dei parametri inseribili è molto più complesso: dal momento in cui l'individuazione dei sospetti consiste semplicemente nel filtrare i residenti in base a valori di determinati campi, si è deciso di lasciare agli utenti la libertà di decidere su quali campi e con quali valori effettuare la ricerca. Per questo motivo, nella form di ricerca viene data la possibilità di im-

postare i filtri su tutti i campi numerici e booleani della tabella *r_redditi*. A fianco della form dei parametri è stato comunque aggiunto un riquadro informativo, in cui vengono specificati i filtri da impostare per replicare le ricerche effettuate nel corso dello sviluppo del pattern. I risultati della ricerca vengono infine rappresentati con il solito formato tabellare; per ogni persona individuata vengono mostrati il nome e il cognome, il reddito totale, il reddito da fabbricati e tutti i campi su cui è stato applicato il filtro (di base non vengono inclusi tutti i campi della tabella *r_redditi* per non rendere troppo complicata la visualizzazione dei risultati).

4.2.3 Il salvataggio dei dati

La possibilità di navigare liberamente il database riconciliato (sia partendo da una ricerca su un residente che dall'esecuzione di un pattern) consente agli utenti di inoltrarsi in percorsi di navigazione potenzialmente infiniti. Nel corso della navigazione, gli utenti potrebbero quindi individuare (più o meno volutamente) soggetti o situazioni che meriterebbero un approfondimento. Tuttavia, tali approfondimenti non possono essere sempre fatti immediatamente: l'utente potrebbe non voler rischiare di perdere di vista la ricerca iniziale che stava facendo, o semplicemente non ha il tempo materiale per guardarci sul momento. Si è pertanto deciso di introdurre una funzionalità di salvataggio delle istanze incontrate durante la navigazione, in modo da averle a disposizione in un apposita finestra e poterle visualizzare ed approfondire in un secondo momento. Ad ogni concetto visualizzabile (residenti, indirizzi, UIU, ecc.) è stato aggiunto un tab, denominato "Salva", in cui viene dato all'utente la possibilità di salvare l'istanza visualizzata. Inoltre, nel tab viene messa a disposizione una form in cui è possibile associare al salvataggio una nota (composta da un titolo e da una generica area di testo): l'utente può così registrare i motivi per cui ha voluto salvare l'istanza, in modo da avere un promemoria nel momento in cui tornerà a visualizzare la suddetta istanza. Per la memorizzazione dei salvataggi e delle note è stato necessario creare una nuova tabella, *int_salvataggi*, nella quale vengono salvate le seguenti informazioni:

- Lo username dell'utente che ha creato il salvataggio.
- Il tipo di istanza salvata (una persona, una UIU, un indirizzo, ecc.).
- L'identificatore dell'istanza salvata (che, a seconda del tipo di istanza, corrisponderà al *cod_master*, all'*id_immobile*, ecc.).

- Il titolo ed il testo della nota associata all'istanza.

La consultazione delle istanze salvate è accessibile attraverso la voce "Salvataggi" nella barra del menù. La schermata di visualizzazione presenta il classico layout a tab utilizzato nella navigazione del database, dove in ogni tab sono elencate le istanze di un determinato concetto. L'elenco delle istanze salvate presenta invece la seguente struttura:

Tabella 13 - Formato dei salvataggi mostrati

Campo	Descrizione
'APRI'	Link per avviare la navigazione sull'istanza.
'CANCELLA'	Link per eliminare il salvataggio dell'istanza.
Descrizione	Informazioni minimali sull'istanza salvata (ad esempio, il nome e il cognome di un residente).
Titolo	Il titolo della nota.
Nota	Il testo della nota.

In conclusione, è importante sottolineare il fatto che si sia voluto mantenere privati i salvataggi degli utenti; ciò significa che un utente è in grado di visualizzare solamente i salvataggi da lui stesso effettuati. Questa scelta è stata adottata principalmente per motivi di semplicità. Tuttavia, la possibilità di condividere i salvataggi o di lasciarli direttamente pubblici deve essere valutata in eventuali sviluppi futuri: qualora un utente dovesse continuare le indagini svolte da un collega, l'accesso ai salvataggi effettuati da quest'ultimo potrebbe sicuramente semplificare il lavoro.

4.2.4 Gestione manuale dei match approssimati

Per completare lo sviluppo dell'interfaccia si è deciso di introdurre la possibilità per gli utenti di intervenire manualmente sui match individuati automaticamente dagli algoritmi di join approssimato. L'obiettivo è quello di far sì che, nel corso della navigazione, gli utenti possano dare una conferma sui singoli match incontrati, accertandone o negandone la correttezza; nel caso si rilevi un errore, gli utenti devono inoltre avere la possibilità di impostare manualmente il match ritenuto corretto. Per realizzare questa funzionalità è stato innanzitutto necessario modifi-

care le tabelle dei match approssimati, aggiungendo un campo che consenta di gestire gli interventi manuali degli utenti. Tale campo, denominato *livello*, specifica il livello di sicurezza del match; i suoi possibili valori sono i seguenti:

Tabella 14 - Decodifica dei valori utilizzati per descrivere il livello dei match

Valore/i	Descrizione
P	Match perfetto, individuato dall'algoritmo di match.
A1, A2, A3	Match approssimato (con un livello di sicurezza decrescente), individuato dall'algoritmo di match.
C	Match confermato: tramite l'interfaccia, un utente ha selezionato un match già esistente e ne ha accertato la correttezza.
E	Match escluso dall'utente: tramite l'interfaccia, un utente ha selezionato un match già esistente e ne ha dichiarato l'erroneità.
M	Match impostato manualmente: tramite interfaccia, l'utente ha cercato e impostato manualmente il match corretto.

In tutte le tabelle di match, il campo *livello* viene inizializzato con i valori {"P"; "A1"; "A2"; "A3"} in base alle caratteristiche dei singoli match (ovvero la regola con cui sono stati individuati e i valori calcolati per le distanze). I valori {"C"; "E"; "M"} vengono invece inseriti (sovrascrivendo i vecchi valori) in base alle azioni dell'utente.

I match su cui l'utente deve avere la possibilità di intervenire sono quelli sulle persone (titolari-residenti) e sui toponimi (catasto-toponomastica e utenze-toponomastica). Pertanto, le schede relative ai concetti interessati (che corrispondono ai titolari, alle UIU e alle utenze) sono state dotate di un tab aggiuntivo, in cui i match coi residenti o coi toponimi possano essere appositamente gestiti. Di seguito viene riportato un esempio per spiegare il funzionamento del meccanismo di gestione manuale dei match.

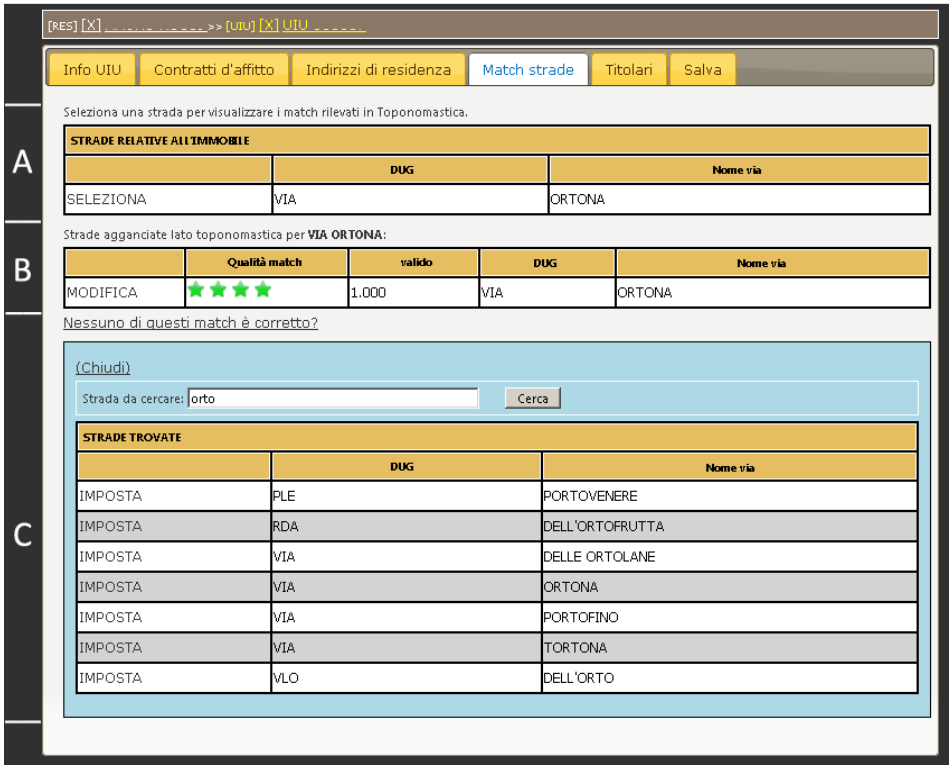

Figura 25 - Screen dell'interfaccia: gestione manuale dei match dei toponimi

Nella figura viene mostrato il tab dedicato alla gestione manuale del match dei toponimi all'interno della scheda di una UIU. Le lettere riportate sul lato sinistro dell'immagine sono state aggiunte per facilitare la descrizione dei componenti:

Tabella 15 - Descrizione dei componenti dell'interfaccia per la gestione manuale dei match

Comp.	Descrizione
A	Si tratta dell'unico pannello visibile all'apertura del tab. Dato il fatto che un'UIU può essere associata ad indirizzi diversi nel corso del tempo, viene chiesto all'utente di selezionare il toponimo per cui si vuole gestire il match. Cliccando su "SELEZIONA" si apre il pannello B.

B	In questa tabella vengono mostrati i match approssimati relativi al toponimo selezionato. Qualora il match per il toponimo sia già stato impostato manualmente, viene mostrato unicamente l'aggancio confermato (il quale non può più essere modificato). In caso contrario viene mostrato l'elenco completo degli agganci individuati; nell'esempio in Figura 25 ne è presente uno solo, ma in altri casi possono essere anche due o più. Cliccando su "MODIFICA" si apre un pop-up che permette di impostare come match corretto il toponimo selezionato (*livello = 'C'*) o di segnalarne l'erroneità (*livello = 'E'*). Se l'utente conferma il match, quest'ultimo viene impostato come l'unico match valido per il toponimo in catasto; in caso contrario, il campo di validità del match viene azzerato.
C	Cliccando sul link con la dicitura "Nessuno di questi match è corretto?", si apre il terzo ed ultimo pannello: se l'utente ritiene che nessuno dei match trovati dall'algoritmo sia quello giusto, egli può eseguire una ricerca sui toponimi in toponomastica per individuare l'associazione corretta. Se individuata, è sufficiente cliccare sul link "IMPOSTA" per associare il toponimo selezionato al toponimo in catasto: il match viene inserito nell'apposita tabella (*livello = 'M'*) e viene impostato come l'unico match valido per il toponimo in catasto.

Lo stesso meccanismo è stato replicato per le schede relative ai consumi e ai titolari. L'unica differenza in queste implementazioni è l'assenza del pannello A, in quanto i campi su cui si basano i match (ossia i toponimi nelle utenze e i dati anagrafici) hanno una cardinalità 1-1 rispetto alle utenze e ai titolari.

4.3 Analisi what-if sull'equità fiscale

Lo studio effettuato in quest'ultima parte esime dalla lotta all'evasione fiscale, ma permette di approfondire le potenzialità informative offerte dal database riconciliato. Per questa parte del progetto si è deciso di realizzare un semplice *proof-of-concept* (POC), ossia un'implementazione non completa dell'idea che si vuole sviluppare, ma sufficientemente approfondita da poterne dimostrare la fattibilità. In generale, l'obiettivo fissato è quello di stimare il carico fiscale che grava sulle famiglie, rapportarlo al patrimonio complessivo della famiglia e valutare se il carico fiscale pesato sia equo tra le diverse categorie di famiglie. Le tasse prese in consi-

derazione per questa analisi sono l'addizionale IRPEF e l'IMU, importanti fonti di introito per il Comune e sulle quali il Comune stesso ha la possibilità di variare le aliquote. Inoltre, si vuole studiare l'impatto causato dall'eventuale variazione delle suddette aliquote sul gettito del Comune e sull'equità fiscale tra le famiglie. Come misura del patrimonio familiare si è deciso di impiegare la scala di equivalenza OCSE, stabilita dall'Organizzazione per la Cooperazione e lo Sviluppo Economico (OECD in inglese) [OEC13]. La scala OCSE prevede di rapportare il reddito complessivo di una famiglia alla composizione del nucleo familiare, il quale incide sull'effettiva ricchezza di una famiglia; ad esempio, un introito complessivo di 50000€ annuali per un single o per una coppia con due figli comporta un indice di ricchezza ben diverso. La normalizzazione del reddito sulla scala OCSE può essere riassunta col seguente pseudocodice:

```
pesoCapoFamiglia = 1;
pesoAltriAdulti = (totaleAdulti - 1) * 0.7;
pesoMinoriDi14Anni = totaleMinoriDi14Anni * 0.5;
pesoNucleoFamiliare = pesoCapoFamiglia + pesoAltriAdulti + pesoMinoriDi14Anni;
redditoOCSE = redditoTotale / pesoNucleoFamiliare;
```

Figura 26 - Pseudocodice per il calcolo del reddito OCSE

4.3.1 La realizzazione del modello previsionale

La prima operazione necessaria consiste nel suddividere le famiglie di residenti in categorie rappresentative della popolazione. In particolare, è stata individuata la seguente scomposizione (basata sulla semplice composizione del nucleo familiare):

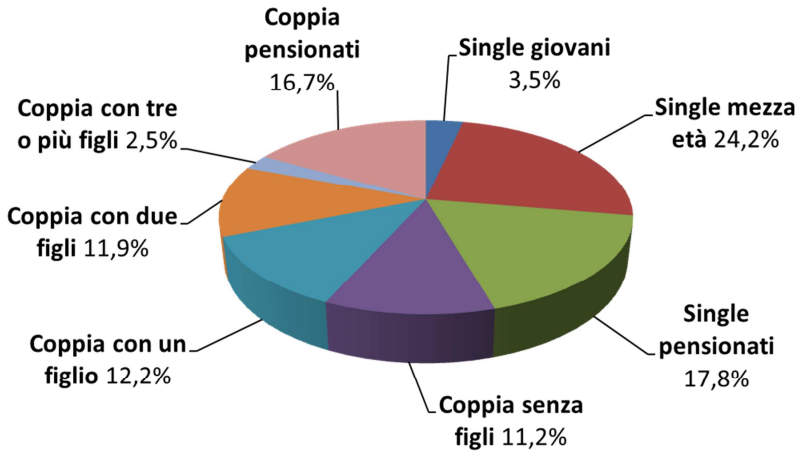

Figura 27 - Suddivisione delle famiglie nel Comune di Cesena

Le categorie di famiglie riportate costituiscono il 75% della popolazione cesenate. Il restante 25% risulta composto da categorie contraddistinte da nuclei familiari più complessi e meno rappresentativi della popolazione, perciò le analisi di equità fiscale sono state svolte sulle sole categorie indicate in Figura 27. Il funzionamento del modello previsionale si basa sul calcolo degli importi dell'addizionale IRPEF e dell'IMU. L'addizionale IRPEF viene calcolata moltiplicando il reddito imponibile di una persona per una determinata aliquota, la quale può variare in base agli scaglioni (fissati dallo Stato per il calcolo dell'IRPEF) in cui rientra il reddito imponibile. In particolare, i parametri necessari per il calcolo sono i seguenti:

Tabella 16 - Parametri per il calcolo dell'addizionale IRPEF

Parametro addizionale IRPEF	Descrizione	Valore fissato dal Comune di Cesena
Soglia di esenzione	Se il reddito imponibile è minore o uguale a questa soglia, il residente è esentato dal pagamento dell'addizionale.	10000€
Aliquote scaglioni	Ad ogni scaglione di reddito imponibile può essere associata un'aliquota diversa (così come avviene per l'IRPEF); gli scaglioni sono i seguenti: • Da 0€ a 15000€ • Da 15001€ a 28000€ • Da 28001€ a 55000€ • Da 55001€ a 75000€ • Da 75000€ in su	0.4% per tutti gli scaglioni
Progressività	Con la modalità non-progressiva, il reddito imponibile viene semplicemente moltiplicato per l'aliquota del corrispondente scaglione. Con la modalità progressiva, il reddito imponibile viene "spezzato" nei vari scaglioni, ciascuno dei quali applica la propria aliquota sulla parte corrispondente di reddito.	Modalità non progressiva

In definitiva, il calcolo dell'addizionale IRPEF non presenta problemi sotto il punto di vista implementativo: le uniche informazioni richieste sono i redditi imponibili dei residenti, disponibili nella tabella *r_redditi.* Il calcolo dell'IMU risulta invece più complesso, specialmente per la necessità di distinguere le abitazioni principali da quelle secondarie. La procedura di calcolo può essere espressa secondo il seguente pseudocodice:

```
totaleIMU = 0;
foreach (immobile in immobiliPosseduti) {
  IMUimmobile = 0;
  /* Si calcola il reddito dell'immobile */
  redditoImmobile = rendita * 1.05 *quotaDiPossesso * giorniDiPossesso/365;
  /* Al reddito viene applicato un coefficiente di moltiplicazione che varia in
  base alla categoria dell'immobile */
  redditoMoltiplicato = redditoImmobile * coefficienteCategoriaImmobile;
  if (isPrimaCasa) {
          IMUimmobile = redditoMoltiplicato * aliquotaPrimaCasa;
          /* Detrazione sulla prima casa di 200€ */
          IMUimmobile -= 200 * quotaDiPossesso;
          /* Detrazione di 50€ per ogni figlio a carico fino a 26 anni */
          IMUimmobile -= 50 * nFigliMinoriDi26 * quotaDiPossesso;
  }
  else if (isAffittato) {
          IMUimmobile = redditoMoltiplicato * aliquotaImmobileAffittato;
  }
  else {
          IMUimmobile = redditoMoltiplicato * aliquotaAltriImmobili;
  }
  totaleIMU += IMUimmobile;
}
```

Figura 28 - Pseudocodice per il calcolo dell'IMU

I parametri su cui Comune ha facoltà di scelta corrispondono alle tre aliquote:

Tabella 17 - Parametri per il calcolo dell'IMU

Parametro IMU	Descrizione	Valore attuale per il Comune di Cesena
Aliquota prima casa	Si applica alla prima casa e agli altri fabbricati ad essa pertinenti (cantina, garage, tettoia).	0.4%

Aliquota immobile affittato	Si applica agli immobili che sono stati affittati.	0.76%
Aliquota altri immobili	Si applica a tutti gli altri immobili.	1.06%

A livello implementativo, il calcolo dell'importo dell'IMU risulta sicuramente più complesso. In particolare, si distinguono due grossi problemi; il primo riguarda le quote di possesso degli immobili: sapendo che molte quote non sono specificate e che molte altre non sono affidabili, risulta impossibile determinare con precisione l'importo totale dell'IMU. Il secondo problema riguarda invece la necessità di individuare le prime case e gli immobili affittati. Innanzitutto, per capire se l'immobile è una prima casa bisogna verificare che il titolare sia residente nello stesso immobile. Tale problematica ricorda quella del match tra coordinate catastali ed indirizzi di residenza (paragrafo 3.2.3); analogamente a quella situazione, l'assenza dei civici interni negli indirizzi catastali impedisce di individuare tutti gli agganci. Inoltre, le note problematiche sui contratti d'affitto non permettono di individuare gli immobili soggetti a locazione, per cui essi verranno riconosciuti come "altri immobili". La somma di tutti questi fattori fa sì che per l'IMU non possa essere calcolato il gettito totale preciso, ma se ne possa fare solamente una stima.

Una volta definiti i meccanismi di calcolo delle imposte e i parametri su cui il Comune ha facoltà di decisione è possibile impostare gli scenari da simulare. In particolare, si è deciso di implementare le seguenti simulazioni:

Tabella 18 - Descrizione delle simulazioni implementate

Simulazione	Descrizione
Simulazione IRPEF	Si calcola l'importo dell'addizionale IRPEF per ogni persona e si determina il gettito totale per il Comune. In questa simulazione vengono considerate tutte le dichiarazioni dei redditi del 2010.

Simulazione IMU	Si calcola l'importo dell'IMU per ogni persona e si determina il gettito totale (da dividersi tra lo Stato e il Comune). In questa simulazione vengono considerate tutte le titolarità catastali assegnate ai residenti ed attualmente valide; inoltre, il gettito versato dalle persone giuridiche non viene considerato
Simulazione CFTM	Si calcola il Carico Fiscale Totale Medio (CFTM) delle categorie di famiglie selezionate. Data la discordanza temporale tra l'addizionale IRPEF (calcolata sui dati del 2010) e l'IMU (calcolata sui dati attuali), in questa simulazione vengono considerate le sole famiglie il cui nucleo familiare non ha subito variazioni dal 2010 ad oggi).
Simulazione Equità	Partendo dalla simulazione precedente, si calcola il CFTM medio per ogni categoria di famiglie e lo si rapporta al reddito OCSE medio della categoria stessa.

Le simulazioni citate sono state implementate sfruttando l'interfaccia web descritta al paragrafo precedente. In particolare, è stata realizzata un'apposita pagina PHP (raggiungibile dalla voce "Equità" nella barra del menù) in cui è possibile inserire i vari parametri per il calcolo delle due imposte e lanciare l'elaborazione - la quale prevede l'esecuzione di tutte le simulazioni citate. Al termine dell'elaborazione, a video vengono visualizzati i gettiti totali calcolati per l'addizionale IRPEF e per l'IMU, mentre nel database vengono memorizzati lo scenario generato (corrispondente alla combinazione dei parametri scelti) e i dati relativi alla simulazione sull'equità. Per la memorizzazione di questi dati, il database è stato arricchito con le seguenti strutture dati.

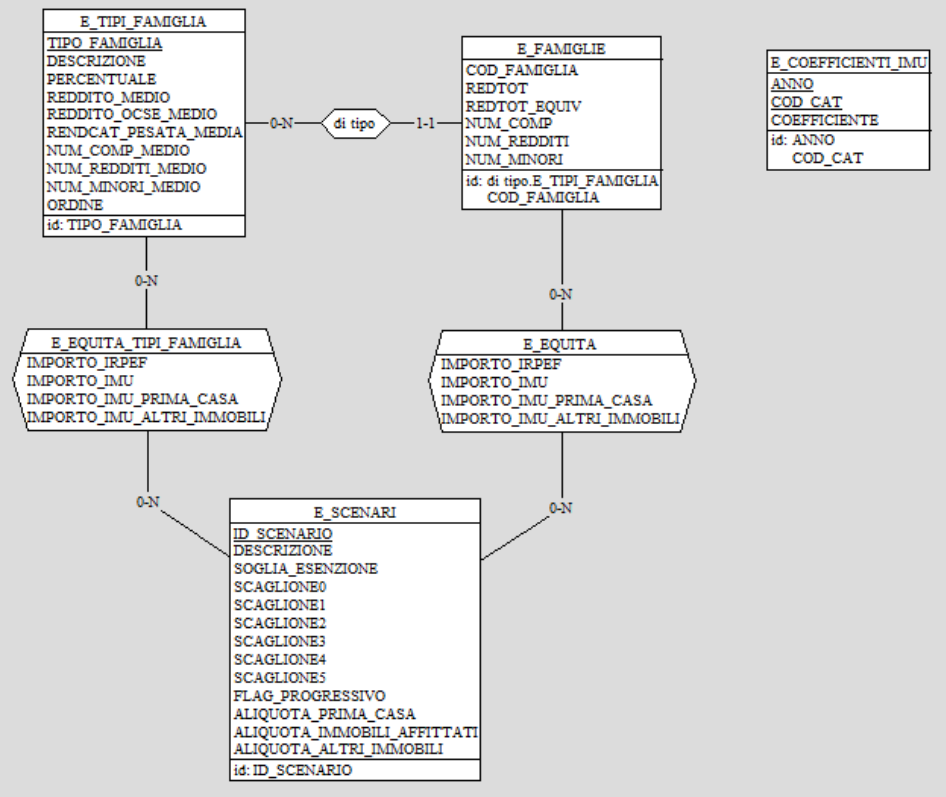

Figura 29 - Diagramma ER delle strutture dati introdotte per il modello previsionale

Il contenuto delle tabelle implementate nel database può essere di seguito riassunto:

Tabella 19 - Descrizione delle tabelle relative al modello previsionale

Tabella	Descrizione
e_coefficienti_imu	Contiene i coefficienti moltiplicativi associati alle categorie catastali degli immobili.
e_famiglie	Contiene i dati delle famiglie su cui vengono eseguite le simulazioni. Il *cod_famiglia* coincide con quello specificato in *a_residenti*.
e_tipi_famiglia	Contiene i dati delle famiglie aggregate per tipologia.
e_scenari	Contiene i parametri degli scenari che si vogliono simulare (ossia i parametri per il calcolo dell'addizionale IRPEF e dell'IMU).
e_equita	Contiene, per ogni famiglia in un determinato scenario, gli importi calcolati dell'addizionale IRPEF e dell'IMU.

e_equita_tipi_famiglia	Contiene, per ogni tipo di famiglia in un determinato scenario, gli importi medi dell'addizionale IRPEF e dell'IMU.

4.3.2 Analisi dei risultati

Il primo passo necessario per un'analisi di tipo previsionale consiste nel conoscere la situazione attuale; pertanto, nel primo scenario simulato sono stati inseriti i parametri applicati attualmente dal Comune di Cesena. Il gettito totale ottenuto per l'addizionale IRPEF (pari a 5.049.739€) risulta congruo con quanto incassato dal Comune nell'anno 2010, perciò viene confermata la correttezza della procedura di calcolo. Per quanto riguarda l'IMU, invece, il gettito totale calcolato per le persone fisiche corrisponde a 33.011.014€, a fronte di un gettito previsto di circa 50 milioni di Euro; se si aggiunge il gettito stimato per le persone giuridiche (calcolato a parte con una query SQL e stimato a circa 5 milioni di Euro) e il gettito proveniente da titolarità di cui non si conoscono le quote di possesso (stimato intorno ai 10 milioni di Euro), i conti sembrano congrui alle aspettative. Una migliore conferma delle procedure di calcolo è stata ottenuta analizzando singolarmente l'IMU calcolata per una dozzina di persone.

Passando alla simulazione del Carico Fiscale Totale Medio delle categorie di famiglie, i risultati ottenuti sono stati disposti nel seguenti grafico:

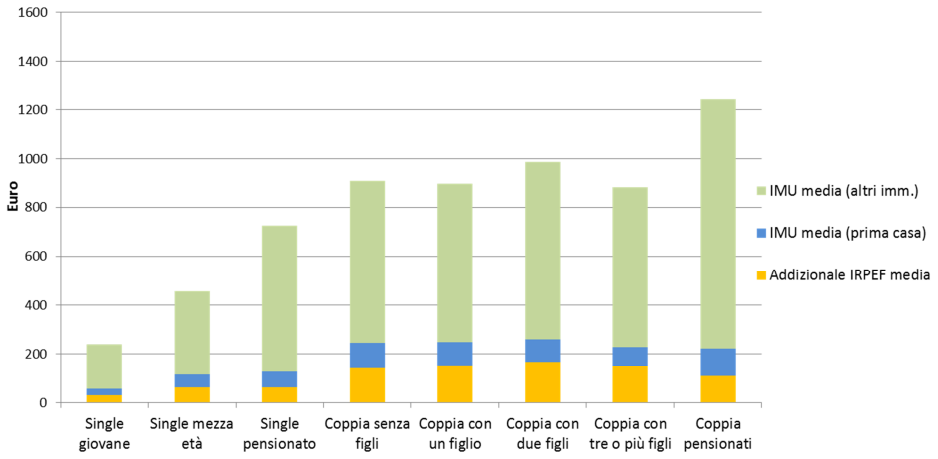

Figura 30 - Carico Fiscale Totale Medio per le famiglie nella situazione attuale

Il grafico in figura mostra una situazione per certi versi prevedibile, con i single giovani che risultano quelli con meno tasse da pagare (causa stipendio tendenzialmente basso e residenza in affitto o in una casa di valore non elevato) e col carico fiscale che raddoppia se si passa dai single di mezza età alle coppie con o senza figli. Un fenomeno interessante è il progressivo aumento dell'IMU corrispondente al crescere dell'età delle persone. I risultati più rilevanti riguardano tuttavia la distribuzione dei rapporti tra i carichi fiscali della famiglie ed i rispettivi redditi OCSE:

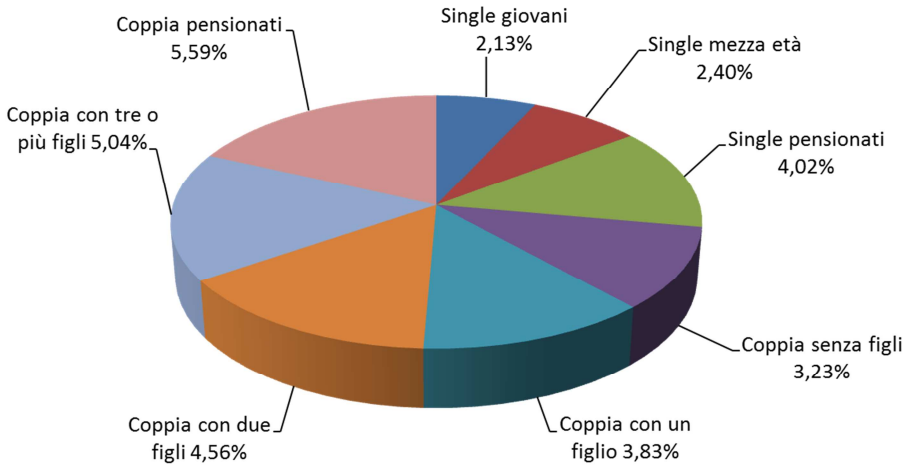

Figura 31 - Rapporto tra il CFTM medio e il reddito OCSE medio della famiglie nella situazione attuale

Come si può notare dal grafico, le coppie di pensionati e le coppie con figli risultano essere le più colpite dalla tassazione: i contributi versati da queste categorie di famiglie incidono sul reddito OCSE circa il doppio rispetto ai single giovani o ai single di mezza età. Una possibile spiegazione potrebbe risiedere nel fatto che il reddito OCSE non tiene conto delle proprietà immobiliari, le quali generano circa l'85% del carico fiscale considerato: sarebbe quindi normale che il peso maggiore delle imposte ricada sulle famiglie che possiedono più beni immobiliari.

Indipendentemente dalle conclusioni che possono essere tratte sulla situazione attuale, si è deciso di simulare una serie di scenari diversi per valutare gli effetti generati dalla variazione dei parametri nel calcolo delle imposte. In particolare, gli scenari simulati sono i seguenti:

Tabella 20 - Descrizione degli scenari simulati con i rispettivi gettiti totali ricavati

Scenario	Gettito totale
IMU prima casa allo 0.5%	34.854.188 €
IMU prima casa allo 0.6%	36.697.363 €
IMU prima casa allo 0.8%	40.383.712 €
Addizionale IRPEF aumentata dello 0.1%	6.313.017 €
Esenzione addizionale IRPEF abbassata a 5000€	5.192.574 €
Addizionale IRPEF progressiva	6.044.036 €

I risultati delle simulazioni in termini di gettiti totali sono indubbiamente interessanti per l'amministrazione comunale, la quale può avere un'indicazione degli incassi previsti a fronte di variazioni ai parametri delle imposte ed essere così in grado di comporre il bilancio in maniera più precisa. Tuttavia, i dati non hanno mostrato particolari evoluzioni dal punto di vista dell'equità fiscale: ad ogni scenario simulato è infatti sempre corrisposto un aumento proporzionale - per tutte le categorie di famiglie - dei rapporti tra il Carico Fiscale Totale Medio ed il reddito OCSE.

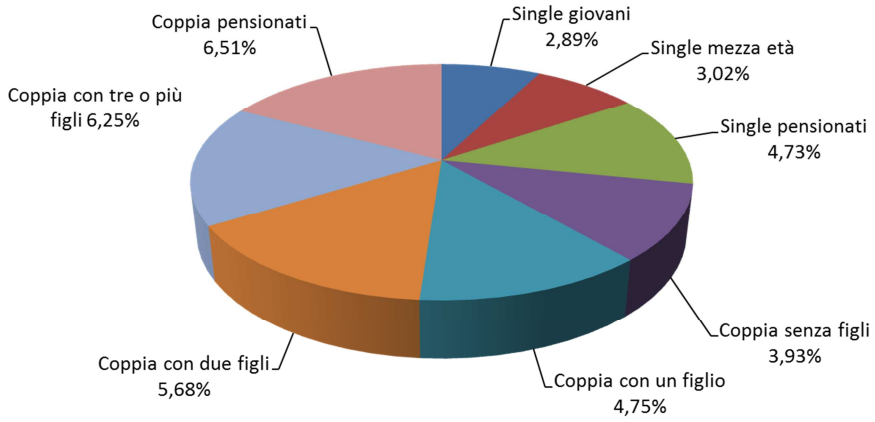

Figura 32 - Rapporto tra il CFTM medio e il reddito OCSE medio della famiglie con l'IMU sulla prima casa allo 0.8%.

La conclusione che si può trarre rispetto a questo fenomeno è che i parametri su cui il Comune ha facoltà di scelta non consentono di intervenire attivamente per migliorare l'equità fiscale tra le famiglie. Questo significa che eventuali strumenti di compensazione della pressione fiscale andranno cercati soprattutto al di fuori dell'ambito tributario, ad esempio introducendo correttivi ai costi dei servizi

erogati tali da ridurre la spesa per le tipologie di famiglie più colpite dalla tassazione.

Conclusioni

L'obiettivo principale di questo progetto è stato quello di sviluppare uno strumento informatico che consenta di individuare situazioni di sospetta evasione fiscale nel territorio di Cesena. Inizialmente si è proceduto con l'analisi e l'integrazione delle banche dati disponibili, dopodiché sono stati definiti ed implementati i pattern per la rilevazione dei casi sospetti. Successivamente è stata sviluppata un'interfaccia web per consentire la navigazione del database integrato l'applicazione dei pattern di evasione. In ultimo è stato implementato un modello previsionale per lo studio dell'equità tra le categorie di famiglie in funzione di diversi scenari fiscali.

I risultati conseguiti al completamento del progetto sono molteplici. Innanzitutto, l'applicazione dei pattern di evasione ha riscosso un esito positivo, raggiunto con l'individuazione di diverse situazioni fortemente sospette e con la segnalazione di alcune di esse all'Agenzia delle Entrate. Sebbene il personale dell'Ufficio Tributi non abbia avuto modo di approfondire tutti i casi presentati, il feedback è stato sufficiente per validare l'attività di ricerca dei potenziali evasori. Inoltre, il fatto che tali risultati siano stati raggiunti nonostante le diverse problematiche riscontrate nei dati provenienti dall'Agenzia delle Entrate costituisce un'ulteriore conferma della bontà delle tecniche adottate; ciò significa anche che, nell'eventualità futura di un miglioramento della qualità dei dati ricevuti, le probabilità di individuare un maggior numero di evasori fiscali non possono che aumentare.

Un ulteriore aspetto emerso nel corso di questo progetto è l'importanza dell'integrazione dei dati sotto diversi punti di vista. In primo luogo, il processo di integrazione ha portato con sé la bonifica dei dati di alcuni concetti comuni;

l'esempio più evidente è quello dei nomi delle strade, in cui le tecniche di join approssimato hanno permesso di aumentare il numero di match tra toponimi dal 55% fino al 95%. Ancora più importante è invece il notevole potenziale informativo offerto dal database riconciliato al di fuori del contesto della lotta all'evasione fiscale. La semplice possibilità di navigare le informazioni integrate costituisce già di per sé un'importante funzionalità: l'interfaccia web sviluppata consente di navigare in maniera fluida ed intuitiva i dati integrati, costituendo un'importante agevolazione per il personale impegnato in attività di controllo. Un'ulteriore dimostrazione è stata concretizzata con la realizzazione del modello previsionale sull'equità fiscale. Sebbene non si sia stata sviluppata una funzionalità completa, l'obiettivo è stato sicuramente raggiunto: non solo si è implementato un possibile utilizzo alternativo del database riconciliato, ma è stato anche provato come i dati integrati possano essere sfruttati efficacemente a scopo predittivo piuttosto che semplicemente descrittivo.

In conclusione, si può affermare che il progetto sia stato portato a termine con successo. Le possibilità di sviluppi futuri restano comunque diverse. Innanzitutto sarebbe utile la realizzazione di una procedura per l'aggiornamento della banca dati: sebbene i processi di caricamento siano stati ben documentati, la loro esecuzione può risultare non immediata e onerosa in termini di tempo; una procedura automatica permetterebbe invece di semplificare l'aggiornamento del database e di risparmiare tempo utile. In secondo luogo sarebbe interessante il perfezionamento dei pattern implementati a seguito dei feedback ricevuti, sia per quanto riguarda i falsi separati, ma soprattutto per quanto riguarda le analisi sui patrimoni economici ed immobiliari: a causa della mancanza di tempo, infatti, il feedback correttivo ricevuto su quest'ultimo pattern non è stato messo in pratica.

Un altro possibile sviluppo futuro è quello di espandere il database riconciliato integrando nuove banche dati, al fine di supportare ulteriormente la lotta all'evasione fiscale o di aprire la porta a nuove funzionalità. Innanzitutto, una delle difficoltà riscontrate nel corso del progetto riguarda il riconoscimento dell'UIU corrispondente alla prima casa, qualora il residente possieda più appartamenti. Per circoscrivere questo problema si potrebbero integrare i dati GIS (Geographical Information System), gestiti dall'Ufficio SIT (Sistema Informativo Territoriale) e

che - tra le altre cose - associano le coordinate catastali agli indirizzi di residenza (ma solo fino al civico esterno). Alternativamente potrebbe essere utile integrare i dati sui pagamenti ICI/IMU (di cui dispone l'Ufficio Tributi), in modo da riconoscere la prima casa attraverso il match sulla rendita catastale. In secondo luogo, un interessante contributo informativo potrebbe essere portato dai dati del PRA (Pubblico Registro Automobilistico, gestito dall'ACI) sulle proprietà dei veicoli, i quali potrebbero essere sfruttati per approfondire le ricerche sulla congruenza tra i redditi dichiarati e i patrimoni posseduti; tuttavia, l'accesso a questi dati non è gratuito, perciò la loro eventuale inclusione andrebbe ben ponderata.

Un'ultima indicazione riguarda la possibilità di completare lo sviluppo del modello previsionale sull'equità fiscale. Innanzitutto potrebbe essere interessante l'applicazione del modello alle persone giuridiche; tale operazione richiederebbe a sua volta un'espansione del database riconciliato per includere le informazioni necessarie. In secondo luogo, il modello potrebbe essere sfruttato per implementare simulazioni più complesse: un esempio potrebbe essere la valutazione della variazione dell'equità fiscale in conseguenza (oltre ai parametri già considerati) di una variazione dei redditi, in particolare di un impoverimento delle categorie di famiglie. Un'altra possibilità riguarda invece l'applicazione di tecniche di clustering per identificare le categorie di famiglie (in base a reddito, stato di famiglia, ecc.), al fine di evitare eventuali disomogeneità causate dalla separazione basata sulla sola composizione del nucleo familiare. Anche il concetto di equità fiscale potrebbe essere migliorato, adottando una misura di ricchezza che - a differenza del reddito OCSE - tenga conto del patrimonio immobiliare di una famiglia; ad esempio si potrebbe tenere conto delle dichiarazioni ISEE (consultabili puntualmente dall'INPS), le quali riportano la consistenza del patrimonio immobiliare. In ultimo, si potrebbe arricchire il concetto di carico fiscale, aggiungendo ulteriori imposte che gravano sul bilancio familiare e che possono essere caratteristiche di diverse categorie di famiglie (come possono esserlo, ad esempio, le rette dell'asilo nido per le coppie con figli piccoli).

Bibliografia

[MDN13] Mozilla Developer Network, http://developer.mozilla.org/en-US/docs/JavaScript, 2013

[OEC13] OECD, “What are equivalence scales?“, http://www.oecd.org/social/family/35411111.pdf, 2013

[ORA13a] Oracle Corporation, “Market Share“, http://www.mysql.com/why-mysql/marketshare, 2013

[ORA13b] Oracle Corporation, “Operating Systems Supported by MySQL Community Server“, http://dev.mysql.com/doc/refman/5.5/en/supported-os.html, 2013

[PAV12] Pavllo Andi, “Progettazione del database integrato per la lotta digitale all'evasione fiscale”, 2012

[PEN13] Pentaho, http://kettle.pentaho.com/, 2013

[SOR12] Sordoni Nicolò, “Progettazione del database integrato per la lotta digitale all'evasione fiscale“, 2012

[TJF13] The jQuery Foundation, http://jquery.com/, 2013

[W3T13a] W3Techs, “Usage of server-side programming languages for websites“, http://w3techs.com/technologies/overview/programming_language/all, 2013

[W3T13b] W3Techs, “Usage of JavaScript libraries for websites“, http://w3techs.com/technologies/overview/javascript_library/all, 2013

[WIK13a] Wikipedia, “MySQL“, http://en.wikipedia.org/wiki/MySQL, 2013

[WIK13b] Wikipedia, “Levenshtein distance“, http://en.wikipedia.org/wiki/Levenshtein_distance, 2013