POLITECNICO DI TORINO

Corso di Laurea Magistrale
in Ingegneria Matematica

Tesi di Laurea Magistrale

Modelli previsionali di produzione di rifiuti nel territorio

Relatori
Gianluca Mastrantonio
Edoardo Fadda

Candidato
Valentina Avigliano

Anno Accademico 2020-2021

A chi ha creduto in me

Sommario

Creare e consolidare un'economia circolare che favorisca la raccolta, la trasformazione e il riutilizzo delle materie già a nostra disposizione è la chiave per frenare il cambiamento climatico.

I cittadini sono sempre più interessati a politiche di sostenibilità a salvaguardia dell'ambiente, come evidenzia la crescita dell'utilizzo della raccolta differenziata.

I rifiuti urbani rappresentano il 10% del totale ed ottimizzarne i processi di raccolta e gestione porta ad un miglioramento per l'intera comunità cittadina.

Stimare il tasso di riempimento dei cassonetti permette di decidere quali siti visitare, minimizzando i costi di trasporto e pianificando l'allocazione delle risorse ma assicurando ai cittadini che i cassonetti vengano svuotati prima del raggiungimento di una determinata soglia di accettabilità.

Per fare questo, il principale strumento a nostra disposizione è l'analisi statistica dello storico della quantità raccolta precedentemente nei siti.

Questa tesi è in collaborazione con *moltosenso*, azienda specializzata nella ricerca e nell'implementazione di nuove soluzioni ICT all'avanguardia.

Nell'ambito del progetto O.N.D.E.- UWC (Optimization for Networked Data in Environmental Urban Waste Collection) promosso dalla regione Piemonte ed in collaborazione con Cidiu Servizi S.p.A (Centro Intercomunale di Igiene Urbana) operante nell'area metropolitana di Torino, sono stati posti dei rilevatori, chiamati transponder, sui cassonetti nel territorio.

Durante il ciclo di raccolta, il transponder si associa al rilevatore posto sul camion, in grado di misurare il peso del rifiuto presente in ogni cassonetto svuotato.

Questo progetto ha lo scopo di creare uno strumento ad hoc per le aziende coinvolte che permetta, a partire dai dati a disposizione, di stimare la quantità di rifiuto presente in ogni giorno nei cassonetti, tramite l'implementazione di algoritmi predittivi sviluppati in R, ambiente per analisi statistiche computazionali.

Ringraziamenti

Per il prezioso supporto dimostratomi durante la stesura di questo progetto, ringrazio *moltosenso* ed in particolar modo al correlatore Edoardo Fadda. Per aver messo a disposizione i dati necessari alle analisi ringrazio Cidiu Servizi S.p.A., Nord Engineering, Insis S.p.A. Infine, un ringraziamento speciale al relatore Gianluca Mastrantonio, per la piena disponibilità e comprensione manifestate durante l'intero processo di scrittura.

Indice

Elenco delle tabelle

Elenco delle figure

Capitolo 1

Introduzione

1.1 Rilevanza del Problema

Ignorare gli effetti del cambiamento climatico non è più possibile: a causa delle azioni dell'uomo, in tutto il mondo si può assistere all'innalzamento delle temperature a fenomeni atmosferici estremi a causa delle emissioni di gas serra dovuti all'uso di petrolio, gas e carbone . Oggi la Terra è più calda di circa 1.2°C [1] rispetto all'epoca pre-industriale e la quantità di anidride carbonica immessa nell'atmosfera è aumentata del 50%. [2] Numerose sono le iniziative che, a livello globale quanto europeo, sono state promosse per evitare che le temperature, aumentando ancora, vadano a distruggere l'intero ecosistema: tra queste l'Agenda 2030 [3] per lo Sviluppo Sostenibile.

L'Agenda 2030 è un programma d'azione per le persone, il pianeta e la prosperità sottoscritto nel settembre 2015 dai governi dei 193 Paesi membri dell'ONU. Essa ingloba 17 Obiettivi per lo Sviluppo Sostenibile – Sustainable Development Goals, SDGs – in un grande programma d'azione per un totale di 169 "target" o traguardi che rappresentano obiettivi comuni su un insieme di questioni importanti per lo sviluppo, tra i quali il contrasto al cambiamento climatico.

La principale strategia proposta dall'Agenda 2030 contro i cambiamenti climatici è la creazione di un'economia circolare, da cui si stima dipenda il 39% [4] dei tagli preventivi di livelli di CO2.

L'economia circolare è un modello di produzione e di consumo capace di rigenerare al suo interno ogni materiale in precedenza utilizzato, preservando il nostro sistema dai rischi di mancanza di materia prima e scarsità di risorse dovute alla crescita dei consumi. Per azzerare gli sprechi e usare in modo efficace ed efficiente le risorse è necessario prevenire la creazione del rifiuto e parallelamente reintrodurre tutti i materiali - non più considerati di scarto - reinserendoli nuovamente in cicli produttivi e di consumo.

Solo in questo modo l'Europa, che al momento importa 3 volte la quantità di materie prime che esporta [5], riuscirà a diventare energeticamente indipendente, ottimizzando e modernizzando l'intera filiera di raccolta, trasporto, stoccaggio, riciclaggio e rivendita. Così facendo, l'economia circolare porterà non solo a benefici ambientali ma anche alla creazione di nuovi settori economici in sviluppo: è evidente che il primo passo per la costruzione di una solida economia circolare è comprendere come gestire in maniera ottimale i rifiuti già posseduti.

Per queste ragioni l'Unione Europea ha imposto agli Stati membri, di riciclare entro il 2020 il 50% [6]di flussi di materiali prioritari presenti nei rifiuti solidi urbani: la carta, i metalli, la plastica e il vetro e ha posto come obiettivo al 2035 il riciclaggio del 65% dei rifiuti urbani.[7]

Situazione in Italia Grazie anche a provvedimenti come il Piano Nazionale di Ripresa e Resilienza, nato a seguito della pandemia, l'economia circolare italiana detiene ora un primato a livello europeo: secondo il Rapporto 2021 [8] del CEN -Circular Economy Network - la rete per lo sviluppo sostenibile - analizzando i risultati raggiunti nelle aree della produzione, consumo, gestione circolare dei rifiuti, investimenti e occupazione nel riciclo, riparazione e riutilizzo - l'Italia è prima tra le 5 maggiori potenze economiche europee (Francia, Germania, Spagna e Polonia). Inoltre, secondo i dati presentati nel Rapporto ASviS 2020 "L'Italia e gli Obiettivi di sviluppo

sostenibile", l'Italia si colloca al secondo posto, dopo l'Olanda, [9] nel perseguire gli obiettivi del Goal 12 dell'Agenda 2030, dedicato a sostenibilità di produzione e consumo. Inoltre l'indice di circolarità della materia e la percentuale di riciclo dei rifiuti, al 49.8%, si aggirano intorno al target europeo per il 2020 del 50% con il consumo materiale interno per unità di PIL in costante diminuzione: -27.5% rispetto al 2010. Questo aspetto è molto importante: nonostante la crescita economica - il conseguente aumento della richiesta di materie prime - l'Italia è riuscita ad abbassare i consumi proporzionali.

Raccolta rifiuti urbani in Italia Nel 2019 la produzione nazionale di rifiuti urbani si assesta a quasi 30.1 tonnellate, con un lieve calo (-0.3%) rispetto al 2018, manifestandosi in un guadagno di circa 80 mila tonnellate di scarto.

Secondo il Rapporto rifiuti urbani 2020 redatto dall'ISPRA [10] - Istituto superiore per la protezione e la ricerca ambientale - la produzione pro-capite di rifiuto urbano è di 487 kg annui. In Piemonte il costo pro-capite medio annuo è stato di 160.02 €, a seguito di 493.1 kg prodotti pro-capite.

Sempre nel 2019, il 61.3% dei rifiuti urbani è ottenuto tramite raccolta differenziata, con una crescita di 3.1 punti percentuali rispetto al 2018: dal 2008 la raccolta differenziata è raddoppiata, passando da circa 9,9 milioni di tonnellate a 18,5 milioni di tonnellate. La componente principale di raccolta differenziata è rappresentata dalla frazione organica insieme a carta e cartone (60% del totale), mentre la plastica rappresenta solo il 7,8%.

La filiera della carta in Italia La filiera cartaria è un tipico esempio di economia circolare: il 57% [11] della carta complessiva prodotta in Italia proviene da fibre riciclate - oltre 5 milioni di tonnellate secondo i dati del 2019. Nel comparto degli imballaggi il tasso di riciclo, sempre nel 2019, ha raggiunto l'81% dell'immesso al consumo, ben oltre l'obiettivo del 75% di riciclo fissato al 2025 dalla nuova direttiva europea e in linea con l'obiettivo dell'85% per il 2030. Secondo quando emerso dal 26° Rapporto Annuale Comieco [12]- Consorzio Nazionale Recupero e Riciclo degli Imballaggi a base Cellulosica - il tasso di riciclo degli imballaggi cellulosici supera l'87%, centrando gli obiettivi UE con 10 anni di anticipo.

Con una resa pro-capite media di 57,2 kg/abitante-anno, nel 2020 sono stati differenziati complessivamente quasi 3,5 milioni di tonnellate di materiale cellulosico, con un lieve decremento dello 0,6% sull'ultimo anno, effetto diretto delle restrizioni dovute all'emergenza sanitaria, pari a circa 22 mila tonnellate.

Sensori sui cassonetti Secondo i dati presentati dal rapporto "Smart Waste", di Navigant Research, il mercato delle tecnologie e dei servizi per il trattamento intelligente dei rifiuti a livello globale arriverà a valere oltre 42 miliardi di dollari entro il 2023.[13] La gestione intelligente dei rifiuti urbani, o *smart waste management*, promossa dalle realtà cittadine, favorisce la messa in atto di politiche di ottimizzazione basate sull'IoT (Internet of Things). Tra queste, è promosso l'inserimento sui cassonetti di sensori - chiamati *transponder* - in grado di rilevare il peso del rifiuto una volta raccolto. Questa innovazione permetterà di abbattere i consumi e le emissioni di anidride carboniche dovute a non necessarie tratte di raccolta, portando ad una riorganizzazione dei turni in base ai reali bisogni del territorio, intervenendo puntualmente nelle aree di maggiore accumulo.

In base alle quantità di rifiuto predette da algoritmi di AI - Intelligenza Artificiale - per ciascun cassonetto è possibile fornire agli operatori un giro di raccolta ottimizzato, permettendo a tutti i soggetti coinvolti nei ciclo di raccolta di sfruttare i dati in maniera sempre più efficiente, in un'ottima di miglioramento continuo ed integrazione con le nuove tecnologie.

Questo progetto nasce dalla collaborazione di più aziende presenti nel territorio confinante la città metropolitana di Torino che, in un'ottica di ricerca della sostenibilità, hanno collaborato per permettere l'installazione di sensori - i *transponder* - sui cassonetti presenti nei siti di raccolta.

Questi sensori, chiamati anche tag, identificano in modo univoco il cassonetto. Quando il veicolo assegnato al ciclo di raccolta svuota il cassonetto, viene pesata la quantità di rifiuto svuotata e viene associato il tag di appartenenza dello scarto. In questo modo, si attribuisce a ciascun cassonetto - e di conseguenza ad ogni sito di raccolta - l'esatto quantitativo di rifiuto presente. Utilizzare i dati dello storico delle pesate permette di prevedere i totali futuri, in modo da organizzare quotidianamente gli itinerari di raccolta. Per ogni giorno, infatti, verranno svuotati solo i cassonetti il cui peso totale stimato all'interno sarà superiore ad una soglia prefissa, che consideri anche le esigenze dei cittadini.

1.2 I partner del Progetto

Cidiu S.p.A. [14], è l'azienda che cura nel territorio tutti gli aspetti della gestione del ciclo dei rifiuti: raccolta, trattamento, smaltimento, riciclo, recupero di energia, anche attraverso aziende controllate.

Il territorio servito, ad ovest del capoluogo piemontese, comprende i Comuni di Alpignano, Buttigliera Alta, Coazze, Collegno, Druento, Giaveno, Grugliasco, Pianezza, Reano, Rivoli, Rosta, Sangano, San Gillio, Trana, Valgioie, Venaria Reale e Villarbasse, per una popolazione di circa 260.000 residenti.

Ecco la mappa dei comuni appartenenti a Cidiu 1.1.

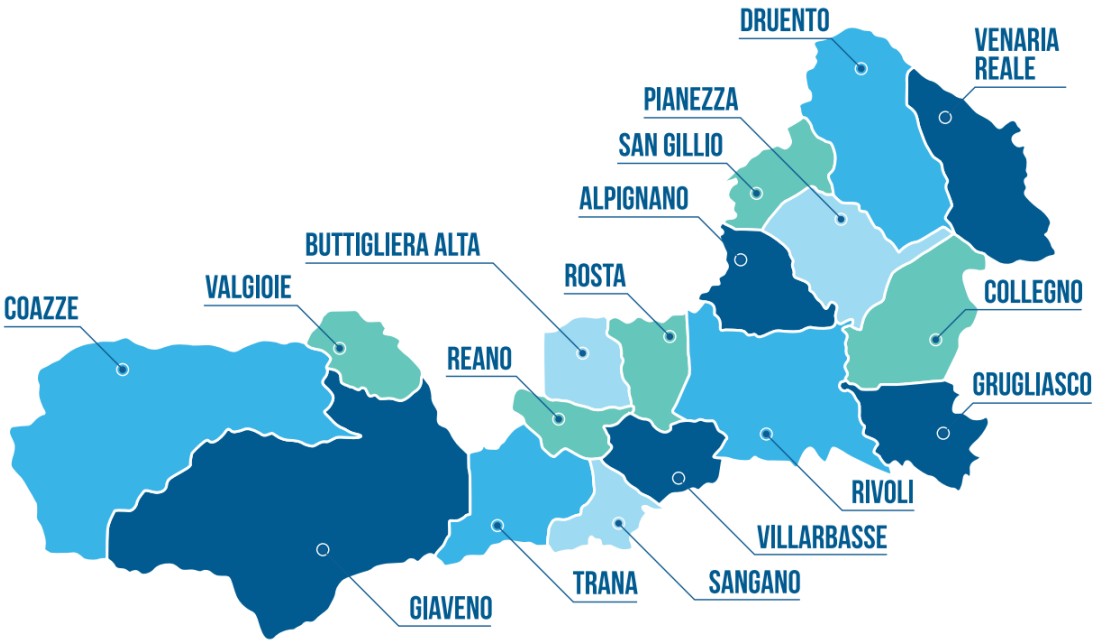

Figura 1.1. Comuni gestiti da a Cidiu

Numerosi transponder sono stati posti nei cassonetti di queste città, permettendo di investire in pratiche di ottimizzazione dei turni di raccolta. Obiettivo dell'azienda è che la totalità dei rifiuti prodotti dai cittadini, anche l'indifferenziato, seguendo differenti percorsi, venga sempre riutilizzato in una nuova forma.

Nord Engineering [15] è tra le aziende leader del settore dello *smart waste management*. Si occupa dello sviluppo di soluzioni altamente tecnologiche, sostenibili ed efficienti realizzate attraverso tecnologie green. Nello specifico, si occupa dell'automazione del ciclo di svuotamento

e riposizionamento dei contenitori altamente funzionali e capaci di contenere da 3.000 a 7.000 litri di rifiuti, sviluppando il sistema di raccolta Easy, capace di raccogliere diverse tipologie di contenitori.

Insis S.p.A. - Fincantieri NexTech - Reicom [16] è una società tecnologica attiva nello sviluppo di soluzioni per la Difesa e la Sicurezza. In un'ottica di ricerca e sviluppo volta alla sostenibilità, ha acquisito Reicom, un'azienda leader nei settori trasporti e telecomunicazioni e si occupa delle pesate i cui dati vengono trasmessi ai server dei partner del progetto.

Il progetto ONDE [17] mette in relazione tra loro le varie competenze che operano nell'ambito dei rifiuti urbani: ha come obiettivo l'utilizzo dell'enorme mole di dati generata dal ciclo della raccolta rifiuti attraverso le informazioni acquisite dai veicoli - veri e propri sensori ambulanti - dai contenitori, dalle banche dati dell'azienda, volto alla totale trasparenza amministrativa. Attraverso Internet of Data, il progetto ha consentito di avvicinare tanto le amministrazioni quanto il cittadino all'azienda che eroga il servizio di gestione dei rifiuti.

moltosenso [18] ha creato tre software applicativi e che rappresentano il cuore del progetto stesso. Il primo è rivolto ai cittadini, il secondo è destinato alle pubbliche amministrazioni e il terzo è pensato per l'azienda erogatrice del servizio di raccolta dei rifiuti. Attraverso la Business Intelligence, gli operatori della raccolta rifiuti possono pianificare i cicli di raccolta in maniera smart e comoda, sia per le istituzioni sia per il cittadino.

Questo elaborato di tesi ha l'obiettivo di analizzare i dati in possesso delle aziende per definire un algoritmo predittivo che integri GLM - Generalized Linear Models - [19] con modelli correlati sia nello spazio che nel tempo. Per fare questo, mi concentrerò sullo storico di pesate dei cassonetti della carta, in quanto di importanza per il territorio - essendo il rifiuto facilmente riciclabile - e poiché si tratta dei dati più puliti e controllati in possesso dell'azienda *moltosenso*. Le medesime considerazioni potranno applicarsi agli altri rifiuti urbani (vetro, rifiuti solidi urbani, plastica). I cassonetti sono organizzati in siti di raccolta localizzati in zone prefissate sul territorio. All'interno di un medesimo sito sono presenti più cassonetti, anche adibiti alla raccolta della medesima tipologia di rifiuto. Il cittadino è libero di gettare i propri rifiuti nel cassonetto che preferisce nel sito in cui si trova: allo stesso modo quando il veicolo per la raccolta giunge in un sito, svuota tutti i cassonetti per tipologia presenti. Per questa ragione, dai dati organizzati per cassonetto, ho ricostruito il database per sito.

Gli aspetti su cui si focalizza l'elaborato sono:

- manipolazione di grandi quantità di dati con controllo dell'attenenza al dominio
- analisi di dati spazialmente correlati
- studio di dati temporalmente correlati ma non equidistanziati: scopo del posizionamento del transponder sul cassonetto è proprio che venga svuotato solo quando necessario, in giorni diversi e ad intervalli irregolari

L'elaborato si articola in 2 parti:

1. Nella prima parte mi focalizzerò sulle fasi di Preprocessing, Pulizia ed Integrazione di dati per permettere di lavorare con osservazioni confrontabili e prive di outlier
2. Nella seconda parte illustrerò gli aspetti teorici relativi all'algoritmo predittivo che ho deciso di sviluppare. Discuterò gli aspetti portanti dell'implementazione dell'algoritmo e concluderò con osservazioni e spunti per ulteriori analisi

Capitolo 2

Metodologia

I dati necessari alle analisi sono stati acquisiti tramite l'interfaccia Python *db_cidiu* fornitomi da *moltosenso*. Le analisi statistiche descritte in questo elaborato sono state effettuate utilizzando l'ambiente R. In particolar modo mi sono servita delle seguenti librerie:

- *sqldf* [20] è un pacchetto per poter utilizzare il linguaggio SQL direttamente nell'ambiente R, usando i dataframe come tabelle, permettendo di velocizzare le operazioni d'integrazione tra molti dati e di T-SQL.
- *ggmap* [21] è un pacchetto che permette, integrandosi con le API di Google Maps, di visualizzare sulla mappa le posizioni geografiche dei luoghi, mostrando le informazioni relative alle vie e città di appartenenza: lo userò per controlli sulla coerenza dei dati
- geosphere [22] permette di calcolare la distanza tra posizioni sulla geosfera, conoscendone latitudine e longitudine
- *arules* [23]permette di generare e visualizzare le regole di associazione tra gli itemset e di determinare gli item che si trovano più frequentemente insieme in una transizione
- *glmmTMB* [24] è un pacchetto per fittare i modelli lineari generalizzzati ad effetti misti (GLMM) tramite approsimazione di Laplace, usando il "Template Model Builder", un tool per implementare complessi modelli ad effetti casuali. Permette di predire modelli che variano per:
  1. distribuzioni di risposta: Poisson, binomiale, binomiale negativa, gamma, beta, normale Gaussiana, t di Student e Tweedie
  2. funzioni di link: log, logit, probit, identità
  3. effetti casuali, anche multipli ed annidati
  4. offset
  5. modelli ad effetti fissi per la dispersione
  6. covarianza degli effetti casuali [25] : diagonale, autoregressiva AR(1), esponenziale, gaussiana, ecc...
- *AICcmodavg* [26] permette di confrontare i modelli predittivi in base a vari criteri: AIC, BIC, log-verosimiglianza, test anova a 2 vie, ecc...

Capitolo 3

Preparazione dei Dati

3.1 Acquisizione dei Dati

L'interfaccia *db_cidiu* richiama i dati presenti nei seguenti database:

- Insis
- Ambiente
- Onde

Ambiente DB è l'interfaccia al database di Ambiente contenente le informazioni anagrafiche di Cidiu, della sua lista cespiti e delle discariche visitate. Le principali informazioni sono:

- anagrafica del contenitore
- anagrafica del servizio preventivato e consuntivato
- anagrafica delle discariche registrate
- elenchi di: targhe dei mezzi di raccolta, comuni con relativi codici numerici identificativi, tipi di rifiuti con codici numerici associati

Come spiegato nella parte introduttiva, sono interessata ad effettuare analisi raggruppando i dati per sito di raccolta - e non per tag - in quanto i cittadini, per ogni materiale, hanno la possibilità di scegliere indistintamente in quale cassonetto gettare i propri scarti e, una volta giunto nel luogo di ritiro, il camion svuota tutti i cassonetti presenti che contengano la giusta categoria di rifiuto. Da AmbienteDb scarico l'anagrafica dei cassonetti con tag attivo, ***anagr***, da cui verranno estrapolati i dati necessari riuniti per sito.

Insis Data Handler è l'interfaccia verso i dati grezzi di Insis, che si occupa di effettuare le pesate e di rendere fruibili nel server i propri dati. Questa interfaccia fa una richiesta al server e di parsifica la risposta in un formato utile per l'interrogazione. Il suo utilizzo permette di verificare i dati presenti in Onde e di avere contezza dei giri di raccolta in tempo reale.

Onde DB contiene i dati forniti da Insis in modo organizzato ed organico. I dati più importanti riguardano:

- i contenitori visitati nel giro: targa, data, ora, tag raccolto, latitudine, longitudine, peso
- i dati sul percorso targa: latitudine, longitudine, stato motore, attivazione della presa di forza e del ribaltamento, velocità
- i dati aggregati: somma dei pesi raccolti, numero di contenitori visitati, ecc.

Su consiglio dell'azienda, per testare la predittività dei modelli, ho scelto di scaricare i dati relativi alle rilevazioni effettuate da gennaio 2020 a fine aprile 2021 ed ho generato due db:

- ***pesate_2020_2021*** che contiene le rilevazioni a cui è stato giustamente associato il tag/cassonetto
- ***errato_sito*** in cui sono presenti le rilevazioni di cui suppongo sia stato individuato giustamente il peso misurato e la posizione geografica ma non il tag del cassonetto svuotato, salvato con il valore di default "00000000AE"

3.2 Descrizione dei Dati Acquisiti

Il database ***pesate_2020_2021*** contiene i seguenti campi:

- X: intero progressivo della rilevazione
- tag: codice alfanumerico che identifica il tag posto sul cassonetto svuotato. Sono presenti 3744 diversi tag attivi ad aprile 2021
- lat e lng: coordinate geografiche (latitudine e longitudine) della posizione del camion durante lo svuotamento del cassonetto
- weight: intero che indica il peso rilevato dall'associazione tra il camion e il transponder/tag
- time: data della rilevazione. Ho deciso di non considerare ora, minuti e secondi ma solamente la giornata
- plate: targa del veicolo che ha svuotato il cassonetto. Sono presenti 18 diverse targhe. Suppongo che il peso del rifiuto svuotato sia indipendente dal camion che ha effettuato l'operazione, ma mi servirò del dato per i controlli relativi all'accuratezza dei dati.

Mostro qui l'header del db:

Listing 3.1. R output

```
##   X        tag      lat      lng       time weight   plate
## 1 0 04190F870D 45.06986 7.559342 2020-01-02     42 FP974BP
## 2 1 04190F5EF4 45.06971 7.556350 2020-01-02     40 FP974BP
## 3 2 160052D37C 45.09566 7.569887 2020-01-02     13 DT513CP
## 4 3 0109A6CED5 45.06941 7.558398 2020-01-02     67 FP974BP
## 5 4 6C008A2FD3 45.09502 7.562747 2020-01-02      5 DT513CP
## 6 5 160052E455 45.09443 7.560928 2020-01-02      8 DT513CP
```

Ogni cassonetto - e quindi ogni tag - può raccogliere un solo tipo di rifiuto. Da questo database non è però possibile comprenderlo né determinare il sito di appartenenza. Per questa ragione è necessario integrare i dati a disposizione con le informazioni presenti nell'anagrafica dei cassonetti.

anagr, l'anagrafica dei tag, contiene i seguenti campi:

- tag : codice alfanumerico per identificare il tag/transponder. Sono presenti 3585 tag attivi a fine aprile 2021 e posizionati giustamente sul territorio
- id_cidiu: identificativo alfanumerico del cassonetto utilizzato dall'azienda Cidiu. Sono stati inseriti solo i cassonetti il cui id_cidiu fosse univoco, per cui sono presenti 3485 diverse valorizzazioni ed esiste una relazione biunivoca con il tag
- id_sito: identificativo alfanumerico del sito di appartenenza. In un sito possono essere presenti più cassonetti, adibiti anche alla raccolta dello stesso tipo di rifiuto: ci sono infatti 1399 siti in cui è presente almeno un tag attivo a fine aprile 2021

- descrizione_sito: descrizione geografica (via e numero civico) della locazione del sito
- comune_cod: identificatore unico numerico del comune di locazione del sito. I tag attivi provengono da 15 comuni situati nella prima cintura della città metropolitana di Torino. I comuni sono i seguenti in tabella 3.1

comune_cod	Nome Comune
000152	Druento
000153	San Gillio
000154	Rivoli
000155	Collegno
000158	Grugliasco
000162	Buttigliera Alta
000277	Rosta
000161	Alpignano
001150	Venaria Reale
000165	Coazze
000279	Trana
000278	Reano
000163	Sangano
000164	Giaveno
000157	Pianezza

Tabella 3.1. Elenco comuni

- latitudine e longitudine: nell'anagrafica, tutti i cassonetti in un sito hanno la medesima posizione geografica. Questi valori sono stati calcolati assegnando ad ogni sito il valore del baricentro delle posizioni dei tag presenti. nella realtà due siti diversi non possono ovviamente avere le medesime coordinate geografiche. Nell'anagrafica a disposizione, a causa di errori di approssimazione delle coordinate, questo può accadere, purché i siti siano situati nella medesima via del medesimo comune, a pochi numeri civici di distanza
- rifiuto : codice numerico per tipo rifiuto raccolto nel tag. Sono presenti 4 materiali: plastica, rifiuti solidi urbani, vetro e carta. Come richiesto, effettuerò le previsioni solamente per il ciclo di raccolta della carta, il cui codice è "200101".

  I rifiuti raccolti con i rispettivi codici sono in tabella 3.2

rifiuto	Nome Rifiuto
150102	plastica
200301	rifiuti solidi urbani
200101	carta
150107	vetro

Tabella 3.2. Elenco rifiuti

- capacita: capacità volumetrica di ogni cassonetto espressa in litri. I cassonetti hanno 6 dimensioni standard, in grado di raccogliere: 2250, 2400, 3000, 3500, 5000 o 7000 litri di scarto. Non è presente il dato relativo alla portata in kg del cassonetto.

- data_ritiro, transponder_al e data_dismissione sono NA perché in questa anagrafica ci sono solo i tag attivi a fine aprile 2021 e, al momento, funzionanti e che non sono stati rimossi. Come concordato dall'azienda, l'analisi verrà svolta solo su questo sotto-insieme di cassonetti poiché non c'è una volontà nell'immediato di creare nuovi siti o di riattivare i tag dismessi
- data_posizionamento : data dalla quale il contenitore è posto in quella locazione. Sono compresi anche cassonetti il cui tag è stato posizionato durante il periodo in analisi
- transponder_dal : data dalla quale è stato inserito il tag sul contenitore. Non è detto che in quella data il contenitore fosse vuoto. Come per data_posizionamento, è un parametro che varia molto, dall'inizio del 2015 ai primi mesi del 2021.

  Mostro qui l'header dell'anagrafica:

Listing 3.2. R output

```
##   tag        id_cidiu        id_sito             comune_cod  latitudine      longitudine
## 1 010A770886 XFH00324        SCAPAPPRD0062103        000155   0.000000        0.000000
## 2 0418110327 XFH00742        SCAPAPPRD0062103        000155   0.000000        0.000000
## 3 04181114E4 XDQ00042        SCAPAPPRD0062103        000155   0.000000        0.000000
## 4 0418112BFF XBH00859        SCAPAPPRD0062103        000155   0.000000        0.000000
## 5 0418113E0C XDQ00043        SCAPAPPRD0062103        000155   0.000000        0.000000
## 6 0418115027 XDQ00044        SCAPAPPRD0062103        000155   0.000000        0.000000

##   rifiuto    capacita   data_ritiro  data_dismissione transponder_al
## 1 150102     2400.000        NA      NA                 NA
## 2 150102     2400.000        NA      NA                 NA
## 3 200301     7000.000        NA      NA                 NA
## 4 150107     2400.000        NA      NA                 NA
## 5 200301     7000.000        NA      NA                 NA
## 6 200301     7000.000        NA      NA                 NA

##   data_posizionamento  transponder_dal
## 1 2018-10-24           2015-03-13
## 2 2018-10-24           2015-10-28
## 3 2020-03-03           2020-03-03
## 4 2018-10-24           2015-10-01
## 5 2020-03-03           2020-03-03
## 6 2020-03-03           2020-03-03
```

L'ultimo database necessario per l'analisi, ***errato_sito***, ha la medesima struttura di ***pesate_2020_2021*** ma il tag è il valore di default "000000AE".

Capitolo 4

Preprocessing Anagrafica

Come detto in precedenza, le analisi verranno svolte per sito e non per tag. Risulta quindi evidente costruire un'anagrafica sito a partire da quella per tag. Il primo passo è pulire il db dei cassonetti, in modo da poter aggregare i valori per sito.

4.1 Pulizia Anagrafica Tag

Passaggio importante prima di procedere all'analisi dati è controllare che tutti i dati presentino valorizzazioni concordi alla definizione del dominio. Per l'anagrafica dei tag ***anagr*** ho controllato che:

- latitudine e longitudine indichino la posizione di punti nelle vicinanze di Torino
- non esistano tag appartenenti a siti diversi ma con le medesime coordinate geografiche, a meno che non siano situati sulla stessa via a pochi numeri civici di distanza
- tutti i tag di un medesimo sito abbiano lo stesso comune_cod

Per prima cosa mostro la descrizione dei campi latitudine e longitudine:

Listing 4.1. R output

```
-------------------------------------------------------------------------------------
latitudine
       n  missing distinct
    3483        4     1292

lowest : 0.000000  45.014433 45.016140 45.017250 45.017891
highest: 45.149710 45.152232 45.155476 45.175410 45.435200
-------------------------------------------------------------------------------------
longitudine
       n  missing distinct
    3483        4     1350

lowest : 0.000000      7.237492      7.244196      7.280350      7.283430
highest: 7.652064      7.652535      7.652890      7.696790      760588.000000
-------------------------------------------------------------------------------------
```

Come si può notare, latitudine e longitudine presentano sia dei valori *missing* che valorizzazioni uguali a 0, inoltre il campo longitudine presenta un valore non ammissibile pari a 760588 che, probabilmente, deriva da un errore di trasmissione dei dati.

Per poter assegnare ai tag la giusta posizione, trasformo questi valori in NA e, poiché in un sito sono presenti più cassonetti, come primo approccio cerco di associare alle valorizzazioni mancanti la posizione degli altri tag posti nel medesimo sito. Infatti, cerco di capire se questa discordanza

tra i dati misurati e quelli trasmessi al db sia dovuta ad errori presenti su un solo cassonetto in un sito

Per tutti i siti con almeno un cassonetto dalla posizione geografica mancante, visualizzo la longitudine e latitudine media degli altri tag presenti.

Listing 4.2. R output

```
##    id_sito           mean_long mean_lat
##    <chr>                 <dbl>    <dbl>
##  1 INTPAPPUB065428          NA      NA
##  2 SCAPAPPRD0062103         NA      NA
##  3 SCAPAPPRD065196          NA      NA
##  4 SCAPAPPRD065214          NA      NA
##  5 SCASTRCOL0024118         NA      45.1
##  6 SCASTRCOL0042543         NA      NA
##  7 SCASTRCOL0043631         NA      NA
##  8 SCASTRCOL065095          NA      NA
##  9 SCASTRCOL065108          NA      NA
## 10 SCASTRCOL065227          NA      NA
## 11 SCASTRCOL065375          NA      NA
```

Come si può notare, quando non è presente in anagrafica la posizione di un tag, questa manca per tutti i cassonetti appartenenti allo stesso sito: l'errore di localizzazione è presente in tutti i tag posti in un sito. Allora, sfruttando il database delle rilevazioni, assegno a questi tag la posizione media rilevata durante i passaggi del camion dei rifiuti nel periodo scelto per l'analisi.

Alla fine, resterà un unico sito che però non è mai stato visitato nel periodo di interesse e quindi, non essendo rilevante per le nostre analisi, elimino dall'anagrafica tutti i dati relativi ai cassonetti lì presenti.

Mostro la nuova descrizione dei campi latitudine e longitudine:

Listing 4.3. R output

```
-------------------------------------------------------------------------------------------
latitudine
       n  missing distinct
    3486        0     1320

lowest : 44.24711 45.01443 45.01614 45.01725 45.01789
highest: 45.14971 45.15223 45.15548 45.17541 45.43520
-------------------------------------------------------------------------------------------
longitudine
       n  missing distinct
    3486        0     1378

lowest : 7.237492 7.244196 7.280350 7.283430 7.284660
highest: 7.652000 7.652064 7.652535 7.652890 7.696790
-------------------------------------------------------------------------------------------
```

Le coordinate geografiche della città di Torino sono: 45.050000 Nord e 7.666667 Est, per cui le posizioni rilevate sembrano plausibili. Per esserne certa, mostro le coordinate nella mappa in Figura 4.1 la mappa: Come si può notare, posizionando i tag sulla mappa, viene ricreata la medesima morfologia dei comuni presenti nella figura 1.1.

All'interno di un sito ovviamente i cassonetti possono avere la medesima posizione geografica. Cerco però di capire se esistono siti diversi i cui tag hanno la medesima locazione. Per farlo, raggruppando per la coppia di coordinate (latitudine, longitudine) mostro quanti siti diversi condividono almeno un cassonetto nella stessa posizione.

Individuate le 6 coppie di siti, cerco di comprendere se si tratti di errori analizzando il comune e la via di appartenenza presente nel campo descrizione_sito.

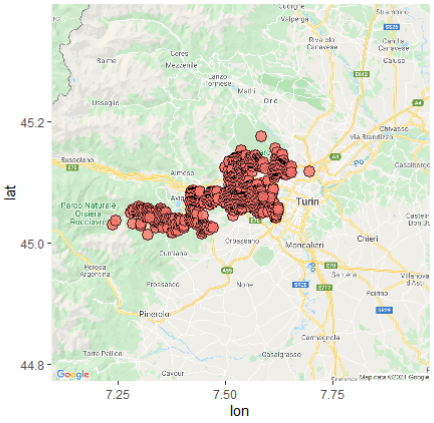

Figura 4.1. Posizione geografica dei tag dopo pulizia dati

Listing 4.4. R output

```
##                id_sito comune_cod                descrizione_sito latitudine
## 1007 SCAPAPPUB0041208      000154              VIA TRIESTE, 00001   45.07202
## 1012 SCAPAPPUB0041209      000154              VIA TRIESTE, 00002   45.07202
## 2137 SCASTRCOL0041595      000158               VIA C.L.N., 00031   45.06725
## 2138 SCASTRCOL0041596      000158               VIA C.L.N., 00040   45.06725
## 2325 SCASTRCOL0040789      000155             VIA FABIO FILZI , 5   45.07241
## 2326 SCAPAPPUB0040788      000155             VIA FABIO FILZI , 2   45.07241
## 2864 SCAPAPPUB0024122      000155                 VIA PASUBIO , 4   45.07294
## 2867 SCASTRCOL0040847      000155                 VIA PASUBIO , 8   45.07294
## 3262  SCASTRCOL065293      001150   Corso Papa Giovanni XXIII, 19   45.12722
## 3264  SCASTRCOL065294      001150 Corso Papa Giovanni XXIII\r\r\n   45.12722
## 3284  SCASTRCOL065425      001150        Via Stefanat Bruno , 103   45.14639
## 3285  SCASTRCOL065423      001150             Via Silva Lelio , 1   45.14639

##      longitudine
## 1007     7.52041
## 1012     7.52041
## 2137     7.56910
## 2138     7.56910
## 2325     7.57362
## 2326     7.57362
## 2864     7.59897
## 2867     7.59897
## 3262     7.63199
## 3264     7.63199
## 3284     7.63378
## 3285     7.63378
```

Come si vede dall'output, è possibile che 2 cassonetti abbiano la stessa locazione se appartenenti alle prime 5 coppie, in quanto situati nella stessa via a pochi numeri civici di distanza. Per quanto concerne invece l'ultima coppia di siti (SCASTRCOL065423 e SCASTRCOL065425) questi si trovano sì nel medesimo comune 0011550-Venaria Reale ma in due vie distanti.

Per comprendere quale dei due siti presenti un errore, mostro sulla mappa le coordinate geografiche e trovo la via: l'errore è relativo al sito SCASTRCOL065423, i cui tag vengono rimossi dall'anagrafica, come si vede dalla mappa in Figura 4.2.

Infine controllo che tutti i tag presenti in un sito abbiano lo stesso comune_cod, raggruppando per id_sito e calcolando quanti comune_cod diversi siano presenti. Poiché il numero massimo e minimo di comuni per un sito è 1, la condizione è verificata.

Mostro in Figura 4.3 le distribuzioni di alcuni campi dell'anagrafica.

Si può notare che solo il 10% dei cassonetti sono adibiti alla raccolta di carta - codice rifiuto "200101" e che i comuni in cui sono presenti più cassonetti sono Rivoli ("000154") e Collegno

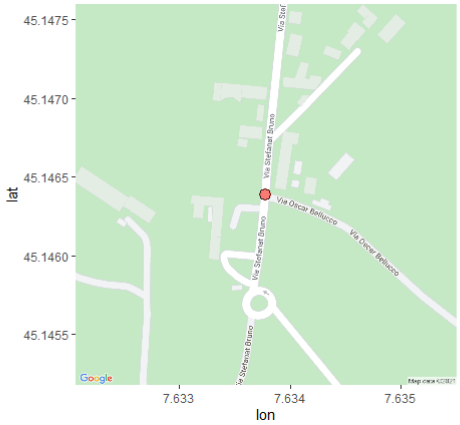

Figura 4.2. Posizione geografica dei tag associati a più siti

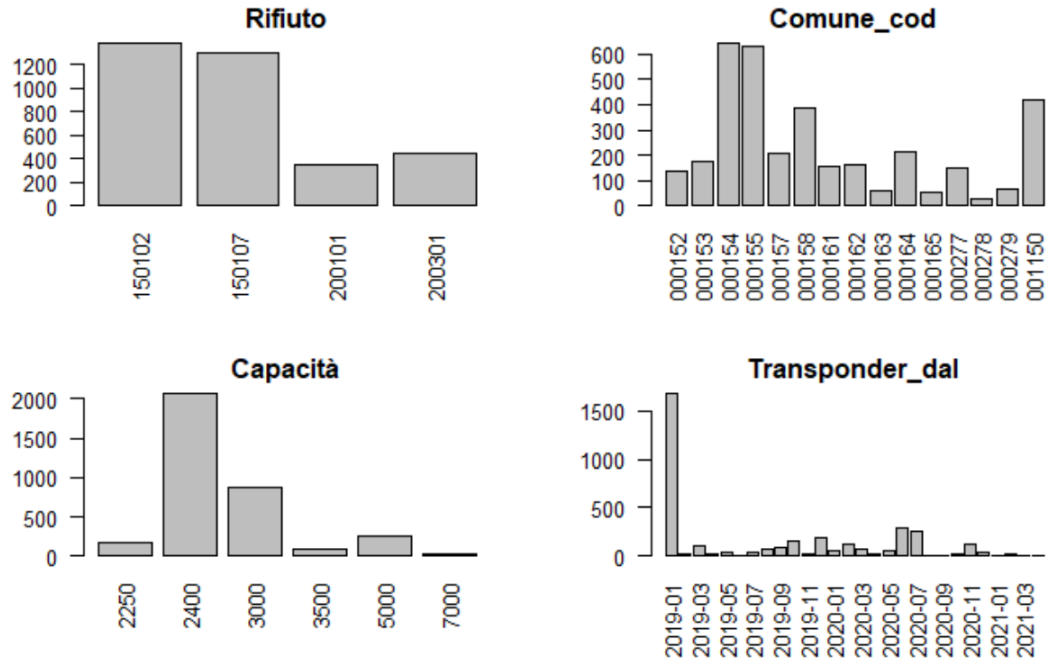

Figura 4.3. Distribuzione dei campi dell'anagrafica tag

("000155"). La maggior parte dei cassonetti sono stai posizionati prima del gennaio 2019, nel cui bin sono raggruppati tutti i posizionamenti antecedenti a tale data.

4.2 Costruzione Anagrafica Sito

A questo punto, pronta l'anagrafica dei tag, creo l'anagrafica dei siti ***anagrafica_sito***, costruita raggruppando ed aggregando per id_sito, comune_cod, descrizione_sito e rifiuto:

- id_sito: sono presenti 1397 siti diversi, dopo aver eliminato sia il sito mai visitato di cui mancavano le coordinate che il sito le cui coordinate geografiche erano erroneamente quelle di un altro
- comune_cod: dai controlli precedenti, è unico per sito. Sono presenti 15 diversi comuni
- descrizione_sito: via e numero civico del sito
- rifiuto: codice per plastica, carta, vetro e rifiuti solidi urbani come da anagrafica dei tag
- tot_tag: numero di tag presenti nel sito adibiti alla raccolta della tipologia di rifiuto selezionata. Ci sono fino a 4 cassonetti dello stesso tipo di rifiuto in un sito
- lat_avg e long_avg: latitudine e longitudine del sito
- pos: data dalla quale è posizionato almeno un tag
- tras: data dalla quale è inserito almeno un transponder su un cassonetto
- capacita: volumi totali che il sito è in grado di accogliere della tipologia di rifiuto selezionata. Va da un minimo di 2250 litri (un cassonetto della dimensione più piccola) a 28000 litri (4 cassonetti della dimensione più grande).

Poiché analizzerò i dati solo per la raccolta della carta, mi concentro sul sotto-insieme dell'anagrafica il cui codice rifiuto è "200101": ***anagrafica_sito_carta***. I siti presenti, 318, sono quelli che contengono almeno un tag attivo ad aprile 2021 per la raccolta del cartaceo

Mostro le distribuzioni dei campi di questa anagrafica in Figura 4.4.

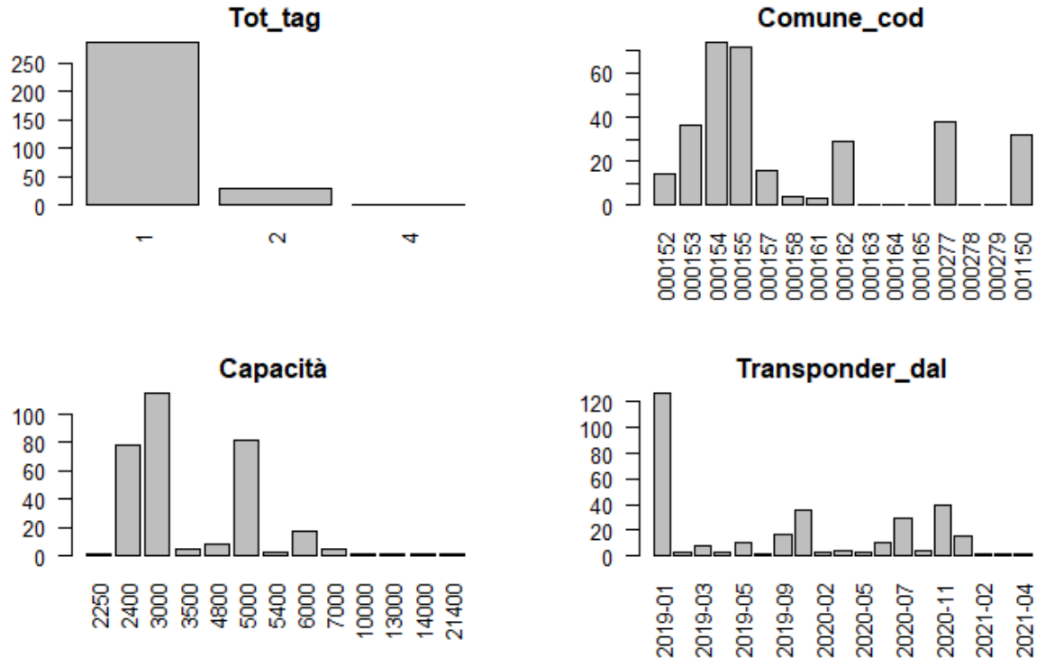

Figura 4.4. Distribuzione dei campi dell'anagrafica sito per rifiuto della carta

Il 90% dei siti ha un solo cassonetto per la raccolta della carta. Il comune che presenta più siti è Rivoli (codice "000154"), seguito da Collegno ("000155").

Anche in questo caso la maggior parte dei siti presenta almeno un cassonetto con transponder da prima di gennaio 2019.

Mostro nella mappa in Figura 4.5 un ingrandimento di quanto presente nella figura 4.1, per i soli siti adibiti alla raccolta della carta nel territorio.

Figura 4.5. Posizione dei siti della carta nel territorio

Come si può notare, questi siti non sono distribuiti uniformemente nel territorio ma anzi, molti di essi si trovano relativamente vicini e sono presenti agglomerati tra le città di Rivoli e Collegno, a Venaria reale e tra Rosta e Buttigliera Alta.

Capitolo 5

Preprocessing Database Misurazioni

Per assicurarmi che i dati presenti del database delle rilevazioni, ***pesate__2020__2021***, siano coerenti, joinando con l'anagrafica dei tag ***anagr***, costruisco un database in cui sono presenti tutte le informazioni derivanti dalle due tabelle e propedeutiche alle analisi dei dati. Controllo che:

- ogni tag sia visitato al massimo una sola volta al giorno
- non siano presenti pesate negative o outlier

Per semplicità, mi concentro solo sul sotto-insieme delle osservazioni relative alla misurazioni della quantità di carta. Nel db ***pesate__2020__2021*** sono presenti 264937 dati. Il sotto-insieme relativo alle solite pesate della carta è composto da 24033 osservazioni, il 9% del totale.

5.1 Pulizia Database Misurazioni

Un tag non può essere stato visitato più volte nelle medesima giornata. Per escludere dal db queste osservazioni, seguo questi passaggi:

1. Raggruppando per tag e giorno, ottengo l'elenco dei cassonetti visitati più di una volta nella medesima giornata: si tratta di 355 osservazioni (1.5% dei dati totali)

2. Tento di capire se la rilevazione sia stata effettuata dallo stesso camion e se quindi si tratti di una - errata - doppia pesata. Per farlo, seleziono le osservazioni che, raggruppate per i campi (tag, plate, time), hanno conteggio superiore ad 1. Ottengo che 314 delle 355 osservazioni precedenti (ovvero l'88% delle volte) finiscono in questa casistica.

3. Cerco di capire se si tratti di osservazioni copiate due volte a database - in cui cambia la localizzazione ma rimane il medesimo "weight" pesato. Raggruppo quindi per la quaterna (tag, plate, time, weight) e seleziono le osservazioni che hanno conteggio superiore a 1. Si tratta di 246 dati sui 314 casi precedenti (78%).

   Allora ripulisco il database di modo che, per ognuna di queste combinazioni weight-plate-time-tag, rimanga una sola osservazione.

4. Per le restanti 68 osservazioni in cui nello stesso giro ci sono state pesature diverse effettuate dallo stesso veicolo, decido di porre come weight totale la somma dei pesi, condensando l'informazione in un'unica osservazione. Un approccio diverso sarebbe potuto essere salvare la media delle rilevazioni o la rilevazione accaduta prima o dopo nel tempo.

5. Quando invece un tag è visitato più volte nello stesso giorno da diversi camion (nei 41 tra i 355 casi iniziali) vedo se è presente un'osservazione con weight = 0 e la elimino, supponendo che, ad esempio, per sbaglio, si sia potuto attivare il meccanismo di associazione tra camion e tag anche quando questo era già stato svuotato. Questo mi permette di trattare 35 tra le 41 osservazioni problematiche: elimino le osservazioni con peso 0.

6. Alla fine rimangono 3 tag che sono stati visitati in due giri diversi nello stesso giorno con entrambe le pesate positive.

   Mostro queste 6 osservazioni nell'output seguente:

Listing 5.1. R output

```
##        X         tag       lat       lng        time
## 1   7883  780095B287  45.06729  7.450737  2020-01-13
## 2   8083  690060517D  45.08447  7.440980  2020-01-13
## 3   8167  690060517D  45.08442  7.441070  2020-01-13
## 4   8273  780095B287  45.06730  7.450820  2020-01-13
## 5 255182  041A4DEFD0  45.07260  7.465470  2021-03-22
## 6 255626  041A4DEFD0  45.07260  7.465590  2021-03-22

##     weight   plate       id_sito            rifiuto   comune_cod
## 1   11       FP974BP     SCAPAPPRD0040596   200101    000162
## 2   141      FP974BP     SCASTRCOL0040631   200101    000162
## 3   10       EA345TP     SCASTRCOL0040631   200101    000162
## 4   14       EA345TP     SCAPAPPRD0040596   200101    000162
## 5   24       GA814FB     SCAPAPPUB0041730   200101    000277
## 6   10       GA813FB     SCAPAPPUB0041730   200101    000277
```

   Si nota che 2 dei 3 tag che presentano questo doppio valore sono stati svuotati il 13 gennaio 2020. Potrebbe esserci stato un disservizio durante uno dei due giri che ha portato a passare effettivamente più volte presso lo stesso sito.

7. Suppongo inizialmente che una delle 2 rilevazioni sia dovuta ad un errore di trasmissione nei dati. Per individuare quale dei due camion sia con maggiore probabilità quello giusto, come primo approccio analizzo quali comuni ha visitato il veicolo in quella giornata. Ovvero, guardo i comuni degli altri siti svuotati dallo stesso camion. Poiché in tutti i casi però gli altri siti appartengono allo stesso comune di questi dati problematici, questo approccio non è utile al fine di rimuovere dei dati errati.

8. Provo allora a vedere se nel sito di appartenenza siano presenti altri cassonetti della carta e da che camion siano stati svuotati in quella giornata problematica: in tutti e 3 i casi si tratta dell'unico cassonetto per il cartaceo nel sito e quindi anche questo approccio non è utile.

9. Poiché nei modelli previsionali non utilizzerò il campo plate come variabile, e ritenendo entrambe le pesate per ciascun tag possibili, elimino dal database la pesata più recente e modifico l'osservazione restante ponendo come weight la somma dei pesi.

5.2 Detenzione degli Outlier

Comprendere se alcuni dei pesi registrati siano degli outlier non è un processo semplice, infatti ci troviamo di fronte a diverse problematiche:

- i cassonetti non sono stati svuotati ad intervalli regolari. Per questa ragione non è facile fare confronti con valori il cui esposto, tempo intercorso dal precedente svuotamento, è molto differente

- i cassonetti si trovano in zone molto diverse del territorio: si tratta sia di quartieri residenziali che di zone industriali e il cartaceo raccolto ha naturalmente caratteristiche diverse (ad

esempio giornali, carta da scrivere ed incarti di cibo nelle zone residenziali, imballaggi di spesso cartonato nelle zone industriali)

- i cassonetti hanno capacità diverse
- a causa delle forti piogge o altri fenomeni atmosferici è possibile che sia entrata dell'acqua nel cassonetto, riempiendolo e bagnando la carta, facendola appesantire

Per effettuare delle prime analisi, mostro la descrizione del campo weight e il valore di alcuni quantili di interesse per comprendere se ci siano outlier :

Listing 5.2. R output

```
       n  missing distinct      Mean       Gmd
   24033        0      348     50.04     44.75

lowest :    0    1    2    3    4, highest: 4879 5001 5020 5863 6593

Percentili

5%  10%  25%  50%  75%
 0    8   22   40   62

90%  91%  92%  93%  94%  95%  96%  97%  98%  99% 100%
 91   94   99  103  108  115  123  135  153  193 6593


99.1%   99.2%   99.3%   99.4%   99.5%   99.6%   99.7%   99.8%   99.9%
200.000 212.000 222.000 234.808 252.520 273.872 324.808 459.680 741.152
```

Come si può notare, il 99% dei dati fa riferimento a pesate inferiori ai 200 kg. Il 99.5% hanno valore inferiore a 250 kg. Mostro in quali giorni sono state effettuate queste rilevazioni problematiche in Figura 5.1

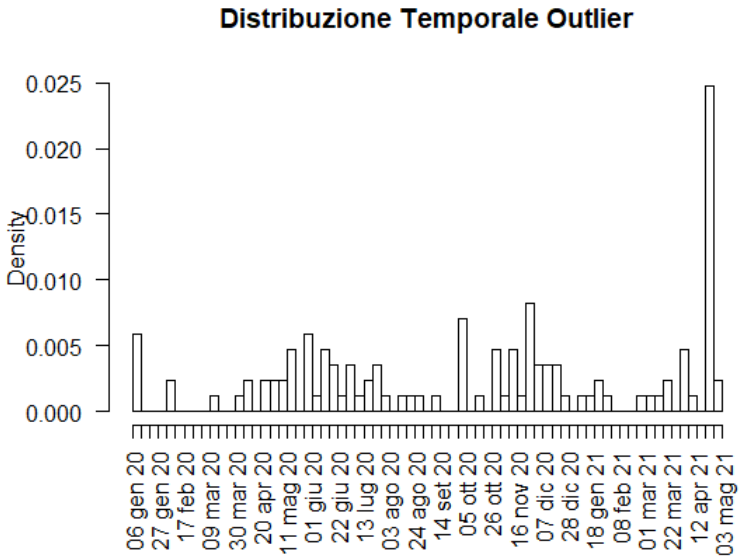

Figura 5.1. Distribuzione temporale delle rilevazioni con peso superiore a 250 chili

Noto che la maggior parte delle misurazioni con valori molto alti sono state effettuate in data 22 aprile 2021, che è anche l'unica giornata in cui sono stati pesati più di 500 kg da un cassonetto. Considero tutte le misurazioni avute in quella data: si tratta di 119 osservazioni, che seguono la distribuzione dei pesi in figura 5.2

Decido di trattare tutte le misurazioni con peso superiore a 250 in due diversi modi:

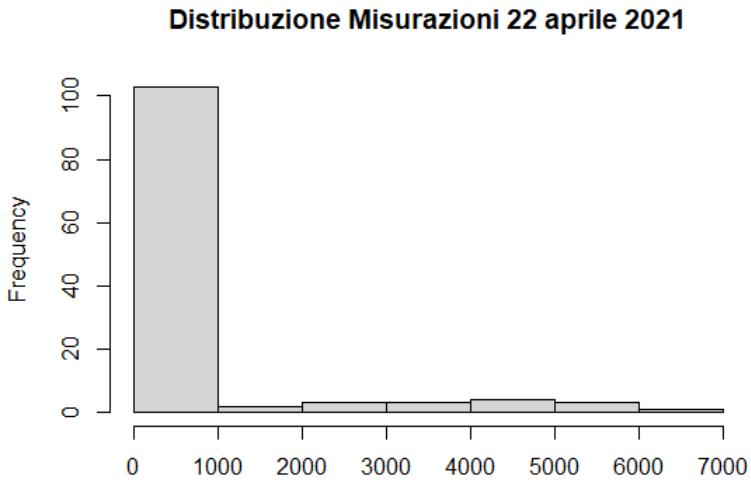

Figura 5.2. Distribuzione delle misurazioni effettuate il 22 aprile

- riportando i valori superiori a 250 kg al cap massimo di 250 per tutti i giorni diversi dal 22 aprile 2021
- eliminando dal db tutte le rilevazioni con peso superiore a 250 se effettuate il 22 aprile 2021: questo perché si tratta di dati presenti nell'ultima settimana del periodo di analisi, che non andrebbero a impattare il modello predittivo temporale

Ho deciso di non considerare le valorizzazioni nulle come valori mancanti, in quanto suppongo che in quella giornata il cassonetto fosse vuoto.

5.3 Pulizia Misurazioni per Sito

A questo punto, dopo aver pulito sia l'anagrafica Tag che il database dello storico delle misurazioni, raggruppo i dati per sito. Controllo inoltre che non solo i tag siano visitati al massimo una volta al giorno, ma che non esistano cassonetti appartenenti ad uno stesso sito ma svuotati da veicoli diversi nella stessa giornata. Mostro nell'output seguente quando questo è avvenuto.

Listing 5.3. R output

```
##        X         tag       lat       lng         time
## 1 182206  6C0085950C  45.0973  7.575620   2020-11-17
## 2 182081  13004821E9  45.0972  7.575508   2020-11-17

##      weight   plate     id_sito             rifiuto   comune_cod
## 1      0      GA814FB   SCAPAPPRD0042134    200101    000155
## 2      0      GB034EB   SCAPAPPRD0042134    200101    000155
```

L'unica volta in cui si è verificata questa problematica, entrambe le misurazioni erano pari a 0. Sommo allora il totale e lo associo al primo tag visitato in giornata.

Capitolo 6

Integrazione del Database delle Errate Assegnazioni

Come detto in precedenza, oltre al dataset ***pesate__2020__2021***, in cui sono presenti le rilevazioni a cui è stato associato il giusto tag, è presente anche un db, ***errato__sito***, con le misurazioni il cui tag è stato erroneamente associato al valore di default "00000000AE". Completo il dataset delle pesate associando alla rilevazione il sito di probabile appartenenza.

1. Questo dataset è composto di 17371 dati per cui, essendo molto vasto, mi concentrerò solamente sulle rilevazioni relative sicuramente alla raccolta di carta per stimarne il sito di appartenenza. Come primo passaggio è quindi necessario capire quale rifiuto sia stato collezionato quel giorno dalla targa, andando ad investigare gli altri cassonetti svuotati dal giro di raccolta.

   Per farlo, estraggo da ***errato__sito*** le combinazioni di (plate, time) che ricerco nel db ***raccolta__sito__carta***.

   A questo punto ho 2773 rilevazioni che sono sicura siano della carta.

2. Poiché per associare il tag mi servirò della localizzazione ottenuta durante lo svuotamento del cassonetto, posso lavorare solo con le osservazioni in cui siano presenti valori possibili di latitudine e longitudine.

   Mostro qui la descrizione dei campi longitudine e latitudine:

Listing 6.1. R output

```
latitudine
       n  missing distinct
    2778        0     1468

lowest :  0.00000 45.02480 45.02850 45.03650 45.04140
highest: 45.14980 45.15547 45.15549 45.15550 45.15553

Value          0    45
Frequency      5  2773
Proportion 0.002 0.998
-------------------------------------------------------------------------------------
longitudine
       n  missing distinct
    2778        0     2081

lowest : 0.000000 7.316730 7.316735 7.316740 7.318600
highest: 7.646637 7.646643 7.646657 7.646680 7.651287

Value        0.0   7.3   7.4   7.5   7.6   7.7
Frequency      5    10    75  1203  1484     1
Proportion 0.002 0.004 0.027 0.433 0.534 0.000
-------------------------------------------------------------------------------------
```

Le uniche geolocalizzazioni che sembrano errate sono le 5 osservazioni hanno latitudine e longitudine pari a 0: elimino questi dati dal db.

3. Per associare il sito di provenienza alla rilevazione calcolo la distanza tra la posizione geografica di ogni sito della carta presente in anagrafica e la rilevazione errata. Utilizzo la distanza Haversine [27], che permette di lavorare con le coordinate geografiche, espressa in metri.

$$D\left(x, y\right) = 2\arcsin\sqrt{\sin^2(\frac{x_1 - y_1}{2}) + \cos x_1 \cos y_1 \sin^2(\frac{x_2 - y_2}{2})} \tag{6.1}$$

dove x_1 e y_1 sono le latitudini e x_2 e y_2 le longitudini dei due punti.

4. Decido di ritenere come possibili siti associabili a una rilevazione solo quelli che distano meno di 150 metri dalla posizione rilevata.

   Calcolo, per ogni rilevazione, quanti siti distano meno di questa distanza di soglia in 6.1:

Listing 6.2. R output

```
conteggio siti
  0   1   2   3   4   5   6   7   8  10  11
969 683 564 284  62  63  97  33  10   2   1
```

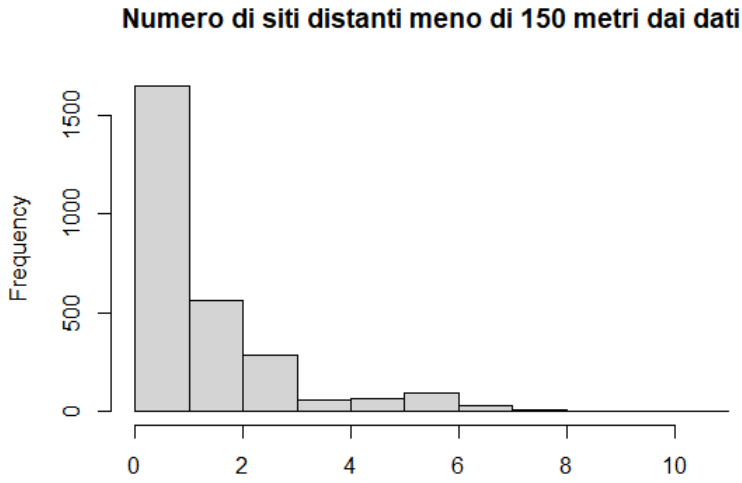

Figura 6.1. Numero siti distanti meno della soglia dalla rilevazione errata

1799 rilevazioni distano meno di 150 metri da almeno un sito, 1116 da almeno due sito, 552 da almeno 3 siti.

5. Non è detto che il sito più prossimo sia effettivamente quello giusto perché potrebbe essere già stato visitato durante quel giro.

   Mi salvo allora le info sui 3 siti più vicini alle rilevazioni in una matrice:

   - idx contiene l'indice (1:n_siti) del sito associabile
   - s contiene la distanza al sito associabile
   - sito contiene l'identificativo alfanumerico del sito associabile

6. Per ogni giorno, elimino dalla lista dei possibili tag associabili, quelli già visitati.

   Associo alla rilevazione il più vicino sito non visitato quel giorno purché distante meno di 150 metri.

In questo modo ho integrato nell'anagrafica 825 nuove osservazioni: posso partire con le analisi propedeutiche alla previsione della raccolta di carta.

Capitolo 7

Analisi Dati

Prima di selezionare l'adeguato algoritmo predittivo, al fine di comprendere meglio l'eterogeneità dei dati, effettuo delle analisi descrittive propedeutiche alla previsione.

Aspetto essenziale del db a disposizione è che i cassonetti non vengono visitati ad intervalli regolari. Nonostante le rilevazioni vadano dal 1 gennaio 2020 al 30 aprile 2021, non in tutti i giorni è avvenuta la raccolta del rifiuto. Dei 485 totali giorni, solo per 361 giorni (ovvero nel 74.4% del totale) è presente almeno un dato.

Mostro il numero totale di rilevazioni per settimana solare in Figura 7.1:

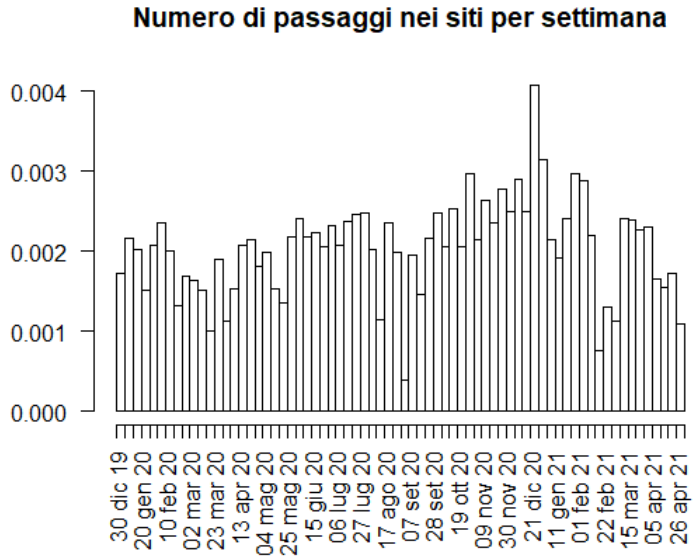

Figura 7.1. Numero di siti visitati per settimana

Come si può notare, il numero di passaggi nei siti è andato aumentando dall'inizio del 2020 fino alla fine dell'anno, per poi calare leggermente.

Una tra le cause delle mancate misurazioni in tutti i giorni in esame, è che non in tutti i giorni della settimana sono stati previsti di cicli di raccolta. Mostro in quali giorni della settimana sono avvenuti i passaggi in Figura 7.2:

Non si effettuano infatti raccolte dei rifiuti della carta la domenica e, tra i giorni lavorativi, il venerdì è il giorno in cui ci sono state meno misurazioni. Questo dato influisce moltissimo sul peso misurato: anche immaginando, per semplicità, che la quantità di rifiuto sia prodotta in modo costante nei giorni, se ad esempio un cassonetto fosse stato svuotato tutti i giorni della

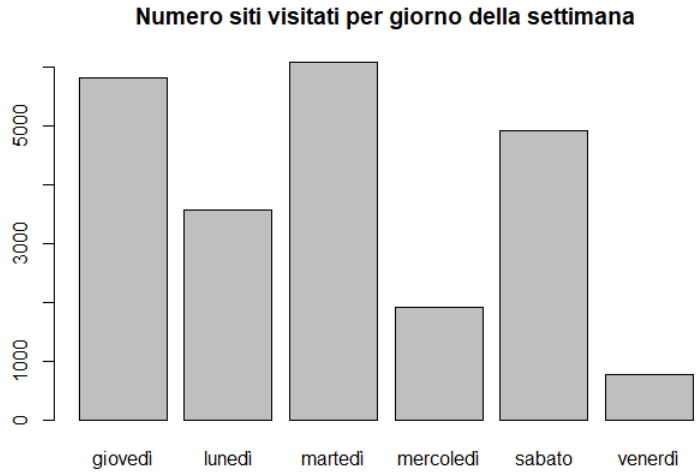

Figura 7.2. Numero di siti visitati per giorno della settimana

settimana tranne la domenica, questo porterebbe ad avere, nelle osservazioni effettuate di lunedì, il doppio dei rifiuti presenti nelle altre giornate.

Immaginando di clusterizzare i dati in base al comune di appartenenza, capire quali comuni siano visitati insieme può fornire delle informazioni importanti per l'analisi. Se 2 siti, di 2 comuni diversi, sono visitati nelle medesime giornate, è possibile determinare una possibile correlazione tra i dati dei due comuni.

Visualizzo in Figura 7.3 quanti comuni sono stati visitati da ogni ciclo di raccolta:

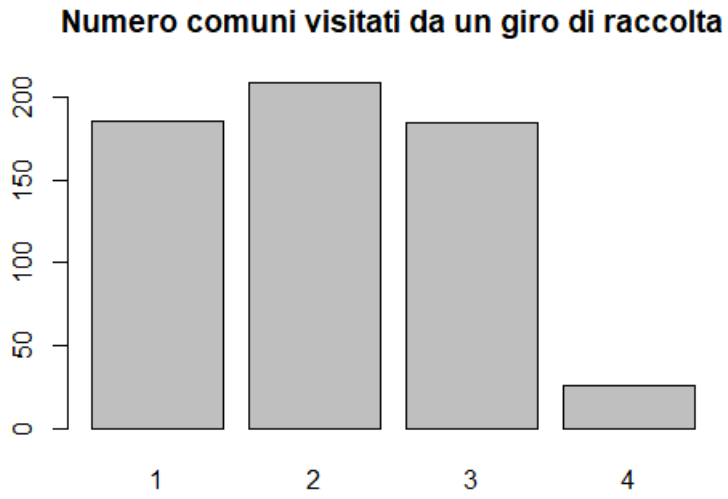

Figura 7.3. Numero di comuni visitati in un giro di raccolta

Sono stati effettuati 607 giri, di cui circa un terzo hanno toccato solo un singolo comune. Mostro quindi, per giorno, il peso medio raccolto da ogni giro 7.4:

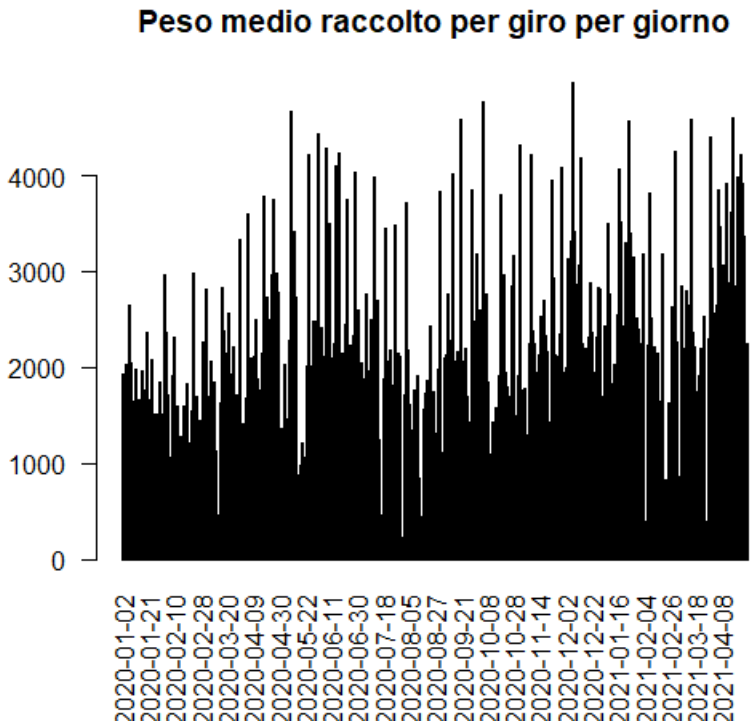

Figura 7.4. Peso totale raccolto giornalmente

7.1 Analisi per Comune

Nonostante in anagrafica siano presenti siti adibiti alla raccolta della carta appartenenti a 15 comuni, nell'intervallo temporale considerato solo 10 di questi hanno fatto parte di almeno un giro di raccolta: Grugliasco, Rivoli, Buttigliera Alta, Rosta, Alpignano, Collegno, Venaria Reale, Druento, San Gillio e Pianezza.

Mostro quanti giri di raccolta hanno riguardato un solo comune, per comune_cod, in Figura 7.5:

A Collegno ("000155") sono presenti numerosi siti i cui giri di raccolta sono stati esclusivamente intra-comune. Ma questo succede anche per i comuni di Rivoli ("000154") e San Gillio ("000153").

Poiché il database lavora su siti appartenenti a diverse zone del territorio, è importante capire le diverse caratteristiche dei comuni considerati.

Creo un dataset dove, per ogni comune visitato, raccolgo delle informazioni utili.

- comune_cod: codice alfanumerico identificativo del comune
- nome: nome del comune
- giorni_sett: in quanti diversi giorni della settimana è stato visitato. Questo dato è utile per distinguere i comuni che sono visitati poche volte, sempre ad intervalli regolari, dagli altri.
- siti: numero totale di siti con almeno un cassonetto adibito alla raccolta della carta presente in anagrafica
- oss_totali: numero di osservazioni totali
- siti_visitati_utili: numero di siti visitati almeno 6 volte
- oss_utili: numero di osservazioni relative a siti considerati utili perché visitati almeno 6 volte

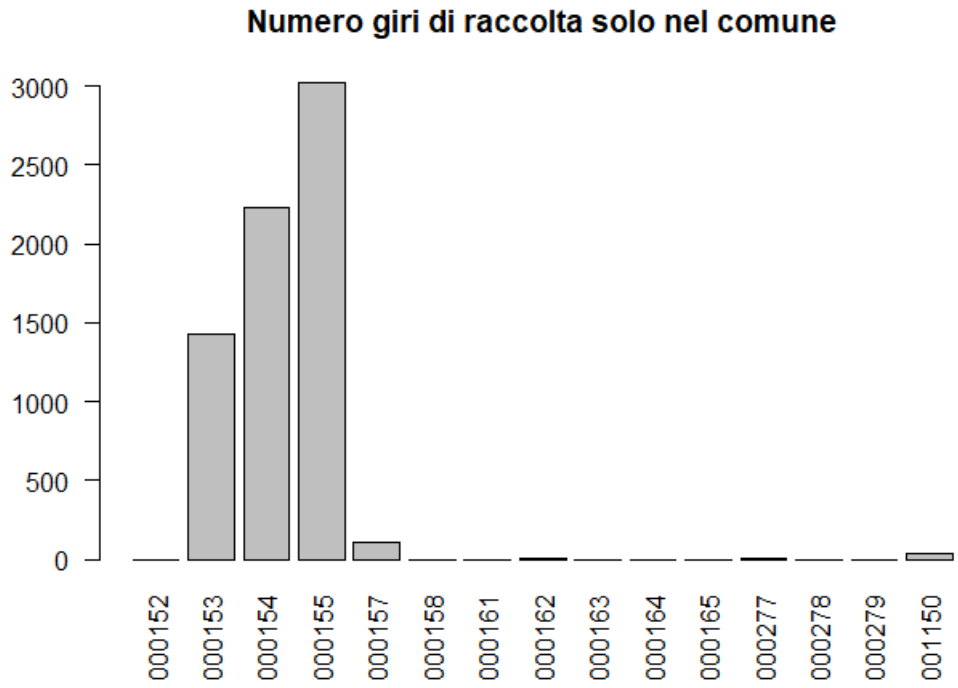

Figura 7.5. Numero di giri di raccolta con un solo comune visitato, per comune

- giorni: in quante giornate diverse un comune è stato visitato
- giri: Poiché i siti di un comune possono essere visitati da giri diversi nella stessa giornata, mostro il totale di giri per comune.

Mostro le informazioni di questi campi per comune:

Listing 7.1. R output

```
##       nome          comune_cod siti oss_totali oss_utili
##  1    Druento          000152   14        245       238
##  2    San Gillio       000153   36       1960      1960
##  3    Rivoli           000154   74       5430      5430
##  4    Collegno         000155   72       5872      5869
##  5    Pianezza         000157   16        965       965
##  6    Grugliasco       000158    4        120       120
##  7    Alpignano        000161    3         86        86
##  8    Buttigliera Alta 000162   29       2878      2878
##  9    Rosta            000277   38       3987      3987
##  10   Venaria Reale    001150   32       1519      1519

##         siti_visitati_utili giorni_sett  giorni giri
##  1                   11          5          66   67
##  2                   36          3          82   82
##  3                   74          6         153  153
##  4                   70          6         199  245
##  5                   16          5         178  189
##  6                    4          4          41   41
##  7                    3          5          72   73
##  8                   29          5         115  116
##  9                   38          4         118  120
##  10                  32          6         166  171
```

Si può vedere che:

- Rivoli e Collegno sono i comuni in cui sono presenti sia più siti che siti utili all'analisi, poiché visitati almeno 6 volte nel periodo
- Rivoli, Collegno e Venaria Reale sono stati visitati in tutti i giorni della settimana esclusa la domenica
- Collegno è il comune visitato in più giorni (199) ed è inoltre quello visitato da più giri di raccolta (245): in questo comune si trovano siti che quotidianamente appartengono a diversi giri di raccolta
- c'è una grande varianza non solo nel numero di siti presenti, ma anche nel numero di giri che li hanno interessati: ad esempio Alpignano, nonostante abbia meno siti di Grugliasco, è stato visitato più frequentemente
- sono presenti molte misurazioni appartenenti al comune Rosta, nonostante abbia approssimativamente lo stesso numero di siti di San Gillio: i cassonetti vengono svuotati con maggiore frequenza

Utilizzando solamente queste informazioni, mi è impossibile capire come e se clusterizzare i comuni in base ai dati di raccolta.

Per capire quali comuni vengano visitati insieme, decido di utilizzare le regole di associazione. [28] Le regole di associazione sono delle tecniche di data mining che permettono di estrarre relazioni nascoste tra i dati, scoprendo quali elementi sono tra di loro associabili in quanto presenti insieme più spesso. Immaginando che ad ogni giro di raccolta corrisponda una transazione, l'idea è stabilire delle regole tra i comuni visitati, che rappresentano gli oggetti nella transazione, non ordinati.

In generale, dato l'insieme di n attributi binari (oggetti o item) $I = \{i_1, i_2, \ldots, i_n\}$ che possono o meno essere presenti in una transizione e l'insieme delle transazioni (database) $D = \{t_1, t_2, \ldots, t_m\}$, una regola è definita come un'implicazione nella forma $X \Rightarrow Y$dove $X, Y \subseteq I$ e $X \cap Y = \emptyset$. Gli insiemi di oggetti (o itemset) X e Y vengono chiamati rispettivamente antecedente e conseguente della regola.

Sia per semplicità la regola di associazione della seguente forma: $X \Rightarrow Y$.

Per valutare la qualità di tutte le regole di associazione, si utilizzano i seguenti indicatori:

- supporto: frazione delle transazioni che includono la coppia (X, Y)
- confidenza: frazione di transazioni che contengono Y sulla totalità delle transazioni che contengono X: è la probabilità condizionata di avere Y trovata X
- lift: confidenza diviso il supporto di X. Non tutte le regole forti sono effettivamente significative: per esserlo, devono verificare anche che l'indicatore di lift sia > 1: questo significa che la regola è più efficace nel predire la probabilità di avere Y nella transazione rispetto alla sua semplice frequenza.

Per generare queste regole, l'algoritmo si basa sul Principio Apriori: ogni sottoinsieme di un itemset frequente deve a sua volta essere frequente. Scelti dei valori minimi di soglia del supporto e della confidenza, si segue un procedimento iterativo a livelli (detto Apriori): estratti gli item più frequenti, si selezionano gli itemset di 2 elementi a partire dagli item più frequenti. Se possibile, si continua aumentando la grandezza degli itemset. Altrimenti, si considerano gli item meno frequenti.

Per applicare questo algoritmo, per ogni coppia (time, plate) che identifica univocamente il giro di raccolta/la transazione, genero l'elenco di tutti i comuni visitati. Come valori di soglia sotto confidenza= 0.4 e supporto= 0.1 e controllo che l'indice di lift sia superiore ad 1.

Ottengo le seguenti 37 regole, ordinate in base alla confidenza decrescente:

Le regole suggeriscono, ad esempio, che:

Listing 7.2. R output

```
    lhs                          rhs       support    confidence coverage   lift     count
[1]  {000158,000161}          => {000154} 0.03150912 1.0000000  0.03150912 3.941176  19
[2]  {000161,001150}          => {000157} 0.01658375 1.0000000  0.01658375 3.190476  10
[3]  {000152,000161}          => {001150} 0.01492537 1.0000000  0.01492537 3.526316   9
[4]  {000152,000157,000161}   => {001150} 0.01492537 1.0000000  0.01492537 3.526316   9
[5]  {000152,000161}          => {000157} 0.01492537 1.0000000  0.01492537 3.190476   9
[6]  {000152,000161,001150}   => {000157} 0.01492537 1.0000000  0.01492537 3.190476   9
[7]  {000158}                 => {000154} 0.06633499 0.9756098  0.06799337 3.845050  40
[8]  {000162}                 => {000277} 0.18739635 0.9741379  0.19237148 4.895043 113
[9]  {000154,000162}          => {000277} 0.06135987 0.9487179  0.06467662 4.767308  37
[10] {000277}                 => {000162} 0.18739635 0.9416667  0.19900498 4.895043 113
[11] {000155,001150}          => {000157} 0.17081260 0.9115044  0.18739635 2.908133 103
[12] {000161,001150}          => {000152} 0.01492537 0.9000000  0.01658375 8.100000   9
[13] {000157,000161,001150}   => {000152} 0.01492537 0.9000000  0.01658375 8.100000   9
[14] {000153,000157}          => {000155} 0.01326700 0.8888889  0.01492537 2.187755   8
[15] {000154,000277}          => {000162} 0.06135987 0.8809524  0.06965174 4.579433  37
[16] {000161,000277}          => {000162} 0.01160862 0.8750000  0.01326700 4.548491   7
[17] {000161,000162}          => {000277} 0.01160862 0.8750000  0.01326700 4.396875   7
[18] {000152}                 => {001150} 0.08955224 0.8059701  0.11111111 2.842105  54
[19] {000157,001150}          => {000155} 0.17081260 0.7984496  0.21393035 1.965164 103
[20] {000155,000157}          => {001150} 0.17081260 0.7744361  0.22056385 2.730906 103
[21] {001150}                 => {000157} 0.21393035 0.7543860  0.28358209 2.406850 129
[22] {000157}                 => {000155} 0.22056385 0.7037037  0.31343284 1.731973 133
[23] {000157}                 => {001150} 0.21393035 0.6825397  0.31343284 2.406850 129
[24] {000152,000157}          => {001150} 0.03648425 0.6666667  0.05472637 2.350877  22
[25] {001150}                 => {000155} 0.18739635 0.6608187  0.28358209 1.626423 113
[26] {000152,000155}          => {000157} 0.02155887 0.6190476  0.03482587 1.975057  13
[27] {000155}                 => {000157} 0.22056385 0.5428571  0.40630182 1.731973 133
[28] {000161}                 => {000154} 0.06301824 0.5205479  0.12106136 2.051571  38
[29] {000154,000161}          => {000158} 0.03150912 0.5000000  0.06301824 7.353659  19
[30] {000152}                 => {000157} 0.05472637 0.4925373  0.11111111 1.571429  33
[31] {000154,000158}          => {000161} 0.03150912 0.4750000  0.06633499 3.923630  19
[32] {000158}                 => {000161} 0.03150912 0.4634146  0.06799337 3.827932  19
[33] {000155}                 => {001150} 0.18739635 0.4612245  0.40630182 1.626423 113
[34] {000152,000155}          => {001150} 0.01492537 0.4285714  0.03482587 1.511278   9
[35] {000153}                 => {000155} 0.05638474 0.4146341  0.13598673 1.020508  34
[36] {000152,000157,001150}   => {000161} 0.01492537 0.4090909  0.03648425 3.379203   9
[37] {000152,001150}          => {000157} 0.03648425 0.4074074  0.08955224 1.299824  22
```

- quando un giro ha toccato sia Alpignano ("000161") che Grugliasco ("000158"), allora ha visitato anche Rivoli ("000154")
- quando un giro ha toccato sia Druento ("000152") che Alpignano ("000161"), allora ha visitato anche Pianezza ("000157")
- nella quasi totalità dei casi, se il veicolo ha svuotato i siti di Grugliasco ("000158"), è passato anche a Rivoli ("000154")
- non è presente la regola di associazione opposta, ovvero con le soglie minime richieste, non è vero che si può considerare molto probabile che, se il veicolo ha svuotato i siti di Rivoli ("000154"), allora è passato anche a Grugliasco ("000158")
- nella quasi totalità dei casi, se il veicolo ha svuotato i siti di Buttigliera Alta ("000162"), è passato anche a Rosta ("000277")
- è molto probabile anche che, se il veicolo ha svuotato i siti di Rosta ("000277"), sia passato anche a Buttigliera Alta ("000162")

Mostro con un grafo i comuni legati da queste regole di associazione, in Figura 7.6: I comuni si possono dividere in 2 cluster: nel primo sono presenti i comuni di Buttigliera Alta ("000162"), Rosta ("000277"), Rivoli ("000154") e Grugliasco ("000158"); nel secondo: Collegno ("000155"), Venaria Reale ("001150"), Druento ("000152"), San Gillio ("000153") e Pianezza "(000157"). Alpignano ("000161") è collegato sia ai comuni del primo che del secondo gruppo: questo perché, a causa si lavori sul territorio, ha cambiato, per un determinato periodo, il proprio giro di raccolta.

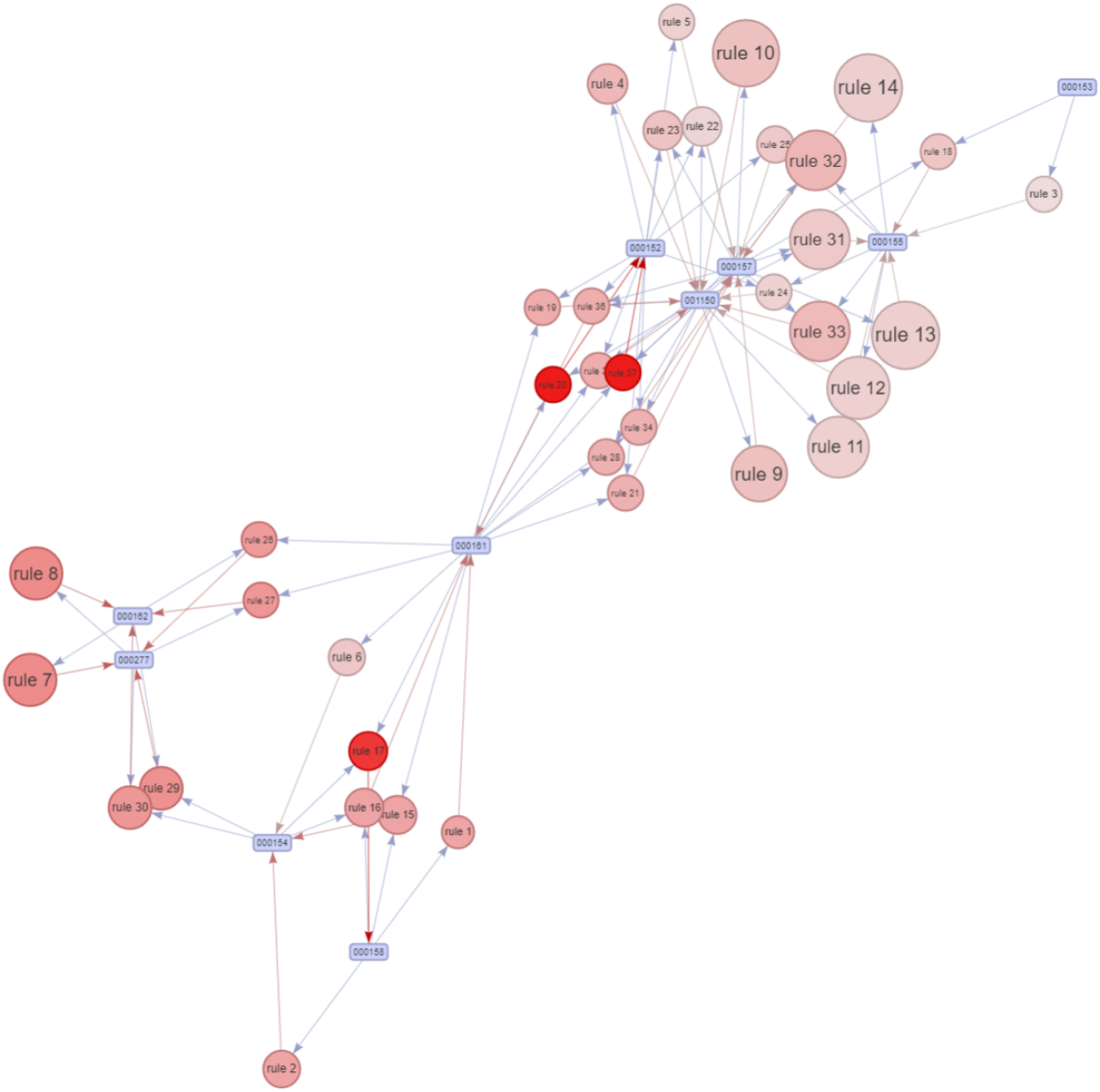

Figura 7.6. Regole di Associazione tra i comuni

Mostro adesso nelle figure 7.7 e 7.8, la quantità di rifiuto raccolta in ogni sito, che è rappresentato da un "pallino".

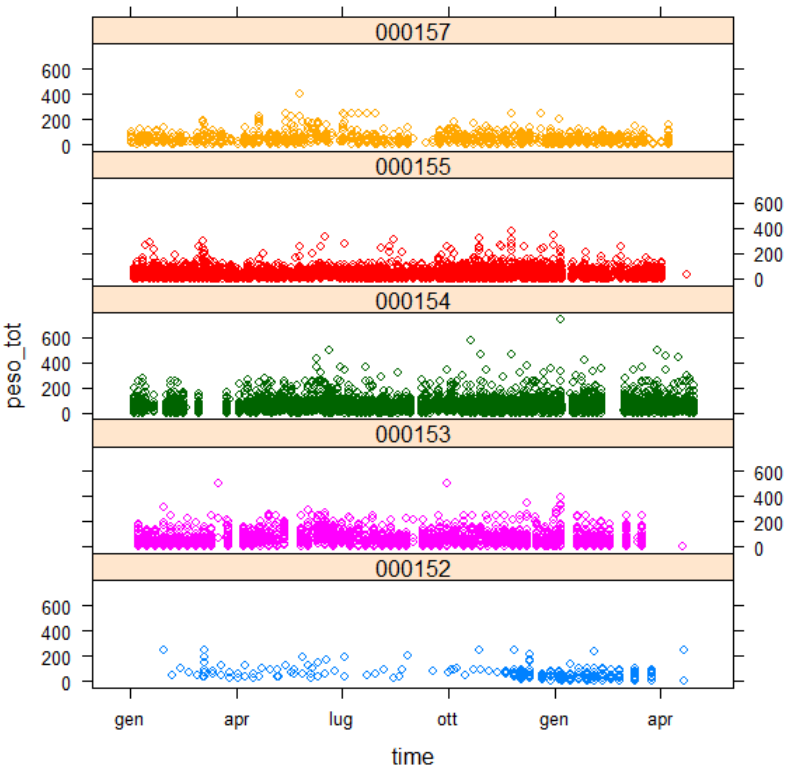

Figura 7.7. Distribuzione temporale dei pesi riscontrati per sito, mostrati per comune

7.2 Analisi per Sito

Analisi generali sui siti.

Creo un dataset dove, per ogni sito visitato, raccolgo delle informazioni utili.

- id_sito: codice alfanumerico identificativo del sito
- giorni_sett: in quanti diversi giorni della settimana è stato visitato. Questo dato è utile per distinguere i siti che sono visitati poche volte, sempre ad intervalli regolari, dagli altri.
- oss_totali: numero di osservazioni totali. Poiché ho richiesto che un sito venisse visitato una sola volta in un giorno, è uguale al numero di giri che lo hanno interessato e quindi al numero di giorni di raccolta in cui il veicolo ha svuotato i cassonetti
- siti_visitati: booleno, vero se il sito è stato visitato almeno 6 volte

Clusterizzare i siti in base a quanto spesso sono stati visitati sembra l'approccio migliore per individuare delle analogie nei dati. Mostro quante volte un sito è stato visitato in Figura 7.9:

La quasi totalità dei siti è stata visitata meno di 120 volte durante i 485 giorni (ovvero meno di 1 giorno su 4) ma è c'è molta varianza tra i valori. Un sito è stato visitato in media circa 73 volte.

Il primo peso rilevato per ogni sito non è utilizzabile per le previsioni in quanto non si conosce l'esposto, ovvero il numero di giorni di accumulo del rifiuto, e non si può da qui dedurre la frequenza/tasso di accumulo. Inoltre, è necessario considerare anche da quanto tempo il sito è in funzione e da quanto tempo è stato posizionato un transponder sui cassonetti.

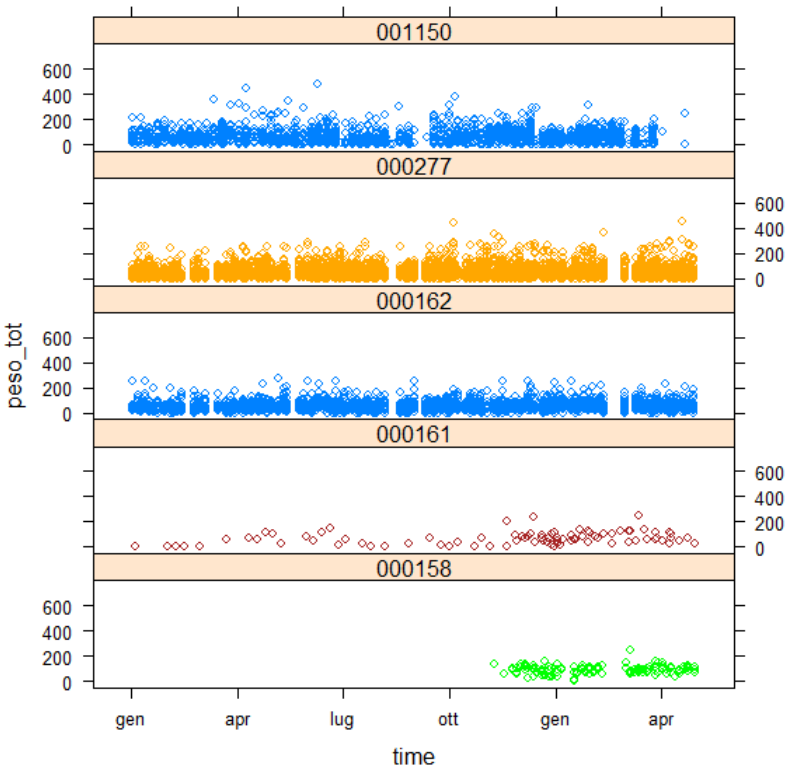

Figura 7.8. Distribuzione temporale dei pesi riscontrati per sito, mostrati per comune

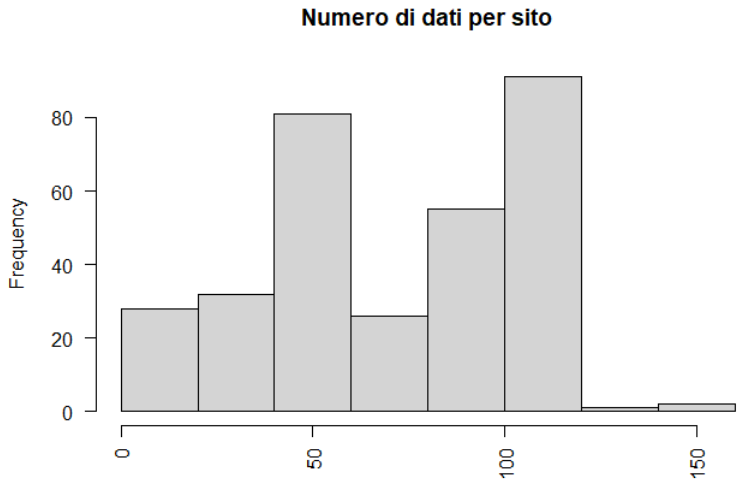

Figura 7.9. Numero di osservazioni per sito

Poiché in ogni sito sono presenti più cassonetti, è possibile che i cassonetti siano stati posti nel sito in date diverse. Allo stesso modo anche i tag possono essere funzionanti da diverso tempo, in quanto posti sul cassonetto in date diverse.

L'ipotesi che faccio è che la quantità di rifiuto prodotta e raccolta per sito non cambi quando si aggiunge un nuovo cassonetto: i cittadini divideranno i propri scarti casualmente tra i cassonetti

presenti.

Capitolo 8

Modelli Predittivi

8.1 Regressione

La regressione è un algoritmo predittivo usato per spiegare la relazione esistente tra una variabile Y - detta variabile risposta oppure output o variabile dipendente - e una o più variabili dette covariate, variabili esplicative, indipendenti, oppure repressori, predittori o variabili di input $(X_1, X_2, \dots X_k)$. [29]

Si ha:

$$Y = f(X_1, X_2, \dots X_k) + \epsilon \tag{8.1}$$

che indica l'esistenza di un legame funzionale in media tra la variabile dipendente e i regressori, rappresentato dalla componente $f(X_1, X_2, \dots X_k)$, denominata componente sistematica. A questa componente va ad aggiungersi un'altra - denominata accidentale, casuale, erronea. Mentre la prima rappresenta la parte della variabile risposta spiegata dai predittori, la seconda componente spiega quella parte di variabilità della risposta che non può ricondursi a fattori sistematici oppure facilmente individuabili, ma dovuti al caso e, più in generale, a cause esterne diverse non prese in considerazione dal modello regressivo.

Il legame funzionale può essere di qualsiasi tipo: la forma più semplice è una funzione di tipo lineare e si parla di regressione lineare multipla o modello lineare che assume la seguente formulazione:

$$Y = \beta_0 + \beta_1 X_1 + \dots \beta_k X_k + \epsilon \tag{8.2}$$

dove β_0 è detto termine noto, mentre $\beta_1, \dots \beta_k$ sono detti coefficienti di regressione e, insieme alla varianza dell'errore, sono i parametri del modello da stimare sulla base delle osservazioni campionarie. In questo caso, si suppone che gli errori seguano la distribuzione gaussiana di media 0 e siano i.i.d:

$$\epsilon \sim \mathcal{N}(0, \sigma^2) \tag{8.3}$$

Diversi modelli, in apparenza non lineari, possono essere linearizzati tramite opportune trasformazioni di variabili.

Quando la variabile risposta non è di tipo continuo si ha una generalizzazione del modello lineare (GLM) che può prendere ad esempio in esame il caso di risposte di tipo dicotomico (regressione logistica) o di conteggio (regressione di Poisson).[30]

I modelli lineari generalizzati sono quindi un'estensione dei modelli lineari, attenuando alcune ipotesi fondamentali nel modello lineare generale, ovvero la linearità del modello di dipendenza, la normalità e l'omoschedasticità delle osservazioni (ovvero la varianza delle osservazioni può non essere costante ed uguale per tutte le variabili).

Famiglia esponenziale di distribuzioni In particolare, l'ipotesi di normalità dei dati del LM (modello lineare) viene generalizzata nei GLM, ipotizzando che la variabile dipendente Y appartenga alla famiglia esponenziale di distribuzioni.

Sia Y una variabile aleatoria discreta o continua la cui distribuzione dipenda dal parametro unidimensionale θ. La distribuzione è detta appartenere alla famiglia esponenziale di distribuzioni se è esprimibile nella forma

$$f(y|\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\} \tag{8.4}$$

Il parametro θ è incognito, mentre il parametro ϕ può essere noto oppure anch'esso incognito. Nel caso in cui ϕ sia noto abbiamo una funzione di densità (o probabilità) della famiglia esponenziale a un parametro (θ). Inoltre a, b e c sono opportune funzioni sufficientemente regolari (continue e differenziabili) rispettivamente di ϕ, di θ e di y e ϕ. Rientrano nella famiglia esponenziale numerose distribuzioni usate nelle applicazioni statistiche: la normale, la binomiale, la Poisson, la tweedie, la beta e la gamma.

Nel caso particolare in cui Y$\sim$ Poisson(μ) con, per semplicità, dimensione n = 1 si ha che:

$$f(y|\mu) = \exp\{y\log(\mu) - \mu - \log(y!)\} \tag{8.5}$$

e

$$\theta = \log(\mu), b(\theta) = \mu = \exp\{\theta\}, a(\phi) = 1, c(y,\phi) = -\log(y!)$$

Un modello lineare generalizzato è individuato da queste caratteristiche:

1. la variabile risposta Y deve una distribuzione appartenente alla famiglia esponenziale
2. le variabili indipendenti influiscono sulla risposta in modo lineare, attraverso il predittore lineare $\eta = X\beta$
3. la media è funzione del predittore lineare, ovvero esiste una funzione g, monotona e differenziabile, detta "link", tale che $E(Y) = \mu = g^{-1}(\eta)$. Compito della funzione link è esplicitare la relazione tra il predittore lineare e il valore atteso della distribuzione.

Se la variabile risposta è una variabile di conteggio (e può assumere di conseguenza solo valori interi non negativi) e le $Y_1, \dots Y_n$ sono n variabili aleatorie poissoniane indipendenti di parametro μ la cui densità assume la forma

$$f(y_i|\mu_i) = \exp\{-\mu_i\}\frac{\mu_i^{y_i}}{y_i!} \tag{8.6}$$

si utilizza la funzione link

$$g(\mu_i) = log(\mu_i) = \eta_i \tag{8.7}$$

in quanto è una funzione che permette di mappare l'intervallo $(0,\infty)$ su tutta la retta reale $(-\infty,\infty)$ e si ottiene

$$log(\mu_i) = \beta_0 + \beta_1 X_{i1} + \dots \beta_k X_{ik} \tag{8.8}$$

o, equivalentemente,

$$\mu_i = exp\{\beta_0 + \beta_1 X_{i1} + \dots \beta_k X_{ik}\} \tag{8.9}$$

infatti

$$g(E(Y_i|x_i)) = \beta_0 + \beta_1 X_{i1} + \dots \beta_k X_{ik} \tag{8.10}$$

Sebbene la scelta di una qualsiasi funzione continua di link g, che va dai numeri positivi all'intera retta reale, possa essere utilizzata per modellizzare variabili poissoniane, nella pratica la funzione prediletta è quella canonica. In questo caso infatti si ha $\theta_i = \log(\mu_i)$ che rappresenta il parametro naturale della legge di Poisson in forma esponenziale e corrisponde al link canonico.

8.2 Regressione di Poisson per Tassi e Frequenze

La distribuzione di Poisson, e quindi il modello regressivo basato su essa, si adatta bene all'analisi e stima di tassi o frequenze. [31] Se il verificarsi di un evento può essere messo in relazione a tempo, spazio o può comunque essere legato ad un indicatore dimensionale misurabile, è preferibile modellizzare il tasso piuttosto che il numero di volte in cui esso si verifica.

Nei casi in cui il confronto dimensionale sia possibile si riscrive il modello in funzione del rapporto tra numero di volte in cui l'evento occorre e la dimensione di riferimento. Sia Y_i la variabile conteggio e sia t_i l'indice dimensionale, si modellizza:

$$tasso_i = \frac{Y_i}{t_i} \tag{8.11}$$

Con il link canonico si riscrive:

$$\log E(\frac{Y_i}{t_i}) = \beta_0 + \beta_1 X_{i1} + \dots \beta_k X_{ik} \tag{8.12}$$

da cui

$$E(\frac{Y_i}{t_i}) = \frac{1}{t_i} E(y_i) = \frac{\mu_i}{t_i} \tag{8.13}$$

che porta a

$$\log \frac{\mu_i}{t_i} = \beta_0 + \beta_1 X_{i1} + \dots \beta_k X_{ik} \tag{8.14}$$

ed infine

$$\log \mu_i = \log t_i + \beta_0 + \beta_1 X_{i1} + \dots \beta_k X_{ik} \tag{8.15}$$

La quantità $\log t_i$ entra nel modello come una variabile esplicativa ma senza parametri associati.

La variabile $\log t_i$ del modello di Poisson per tassi viene detta offset. In generale una variabile offset è una variabile che entra nel modello con effetto noto. Può essere utilizzata nella modellizzazione con scopo correttivo. Si riscrive il modello GLM come

$$g(\mu_i) = \eta_i + OFFSET_i \tag{8.16}$$

e il modello descrive il tasso logaritmico del conteggio di una certa quantità in base all'unità di misura di confronto. Quindi $exp\{\beta_j\}$ rappresenta il cambiamento relativo del tasso per incremento unitario della covariata X_j.

8.3 GLMM

I modelli misti lineari generalizzati (o GLMM) sono un'estensione del modello lineare generalizzato (GLM) in cui il predittore lineare contiene effetti casuali oltre ai soliti effetti fissi. Ereditano anche dai GLM l'idea di estendere i modelli misti lineari a dati non normali.

I GLMM forniscono una vasta gamma di modelli per l'analisi dei dati raggruppati, poiché le differenze tra i gruppi possono essere modellizzate come un effetto casuale. Questi modelli sono utili nell'analisi di molti tipi di dati, inclusi i dati longitudinali. [32]

Sono particolarmente utilizzati quando non vi è indipendenza nei dati, come deriva da una struttura gerarchica: nel nostro caso le osservazioni possono essere raggruppate per siti, a loro volta appartenenti a diversi comuni.

Quando ci sono più livelli, la variabilità del risultato può essere pensata come all'interno del gruppo o tra i gruppi. Ad esempio, nel nostro caso le osservazioni a livello di sito non sono indipendenti, poiché più simili tra loro. Allo stesso modo esiste un ulteriore livello, costituito dai siti presenti in un medesimo comune, ed è possibile valutare come variano i dati all'interno di esso.

Scelto un predittore, possiamo imporre che abbia il medesimo effetto sulle osservazioni di tutto il db oppure che vari effetto all'interno del sito o per tutto un comune.

8.4 Poisson Autoregressivo

Utilizzare un semplice modello di regressione di Poisson quando i dati hanno una forte componente temporale è limitante: infatti le sequenze di dati temporali sono spesso auto-correlate. Risulta infatti naturale pensare che le quantità raccolte nel passato in un sito influenzino i valori futuri. Se il modello di regressione non è in grado di acquisire adeguatamente le informazioni contenute in queste correlazioni, si generano errori residui del modello sotto forma di errori auto-correlati: è necessario introdurre un termine correttivo che sia in grado di cogliere la correlazione temporale. [33] Per farlo, bisogna comprendere come adattare il processo stocastico con termini autoregressivi a tempo continuo.

Si definisce processo stocastico una famiglia di variabili aleatorie $\{Y(t), t \in \mathbb{R}\}$ dipendenti da un parametro t - che in questo caso indica il tempo - definite su uno spazio campione Ξ e che assumono valori in un insieme definito "spazio degli stati" del processo. Fissato l'istante temporale, il processo si riduce ad una variabile aleatoria mentre, fissato l'esperimento ξ, si riduce alla realizzazione del processo. Si può dimostrare che un processo stocastico è completamente caratterizzato da tutte le PDF (funzioni di densità) congiunte delle variabili aleatorie $Y(t_1), Y(t_2), \ldots, Y(t_n)$ ed il processo si definisce stabile se le sue PDF di qualunque ordine sono invarianti rispetto alle traslazioni temporali. Un processo si dice gaussiano se il vettore aleatorio $Y(t_1), Y(t_2), \ldots, Y(t_n)$ segue la distribuzione gaussiana. Il processo è di Wiener (moto browniano) se: parte da 0 quasi certamente, le traiettorie sono continue quasi certamente, ha incrementi indipendenti ed ha incrementi gaussiani di varianza pari alla differenza dei tempi. Il processo si definisce markoviano (catena di Markov) se la probabilità di transizione non dipende dalla totalità della storia pregressa ma solo dallo stato del sistema al tempo immediatamente precedente ovvero non dipende da come si è giunti in quello stato. Un processo si definisce omogeneo se la probabilità di transizione al tempo t_i non dipende dal tempo stesso ma solo dallo stato del sistema al tempo immediatamente precedente $Y(t_{i-1})$. Le serie temporali sono dei processi stocastici.

Un modello autoregressivo AR(p) - dove p è l'ordine del modello - è un processo stocastico lineare che specifica che il valore di una variabile $Y(t_i)$ dipende linearmente dalle valorizzazioni in tempi precedenti, creando autocorrelazione: $Y(t_i)$ è pensato come somma di un termine che dipende dai valori precedenti nel tempo e una parte di novità $\epsilon(t)$, descritta come rumore bianco.

Nel caso AR(1), posto i l'i-esimo sito e j il j-esimo tempo, si riscrive il modello autoregressivo come: [34]

$$y_{ij} = x_{ij}\beta + b_{ij} + \epsilon_{ij} \tag{8.17}$$

con

$$b_{ij} \sim \mathcal{N}(\tilde{\phi} b_{i-1,j}, \sigma_I^2) \tag{8.18}$$

che descrive il processo AR(1) con ϕ la correlazione di lag 1, σ_I^2 varianza dell'innovazione, $\sigma_b^2 = \frac{\sigma_I^2}{1-\tilde{\phi}^2}$ la varianza marginale del processo. Affinché il processo AR(1) possa essere stazionario si impone che $|\phi| < 1$. Inoltre il rumore bianco segue la distribuzione:

$$\epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2) \tag{8.19}$$

Riparametrizzando $\theta_{ij} = b_{ij} + \epsilon_{ij}$, si ottiene, per ogni istante di tempo j:

$$Corr(\theta_{ij}, \theta_{hj}) = \frac{{\sigma_b}^2}{\sigma_b^2 + \sigma_\epsilon^2} \tilde{\phi}^{|i-h|} \tag{8.20}$$

con

$$\theta_j \sim \mathcal{N}(0, \sigma^2 \mathrm{R}_j(\phi)) \tag{8.21}$$

dove $\sigma^2 = \sigma_b^2 + \sigma_\epsilon^2$ è la varianza totale e

$$\mathrm{R}_j(\phi) = \begin{bmatrix} 1 & \phi & \phi^3 & \phi^2 & \cdots & \phi^n \\ \phi & 1 & \phi & \phi^2 & \cdots & \phi^{n-1} \\ \phi^2 & \phi & 1 & \phi & \cdots & \phi^{n-2} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \phi^{n-2} & \cdots & \phi & 1 & \phi & \phi^2 \\ \phi^{n-1} & \phi^{n-2} & \cdots & \phi & 1 & \phi \\ \phi^n & \phi^{n-1} & \cdots & \cdots & \phi & 1 \end{bmatrix} \tag{8.22}$$

dove il parametro di correlazione al generico lag $|i-h|$ è pari a

$$\phi^{|i-h|} = \frac{{\sigma_b}^2}{\sigma_b^2 + \sigma_\epsilon^2} \tilde{\phi}^{|i-h|} \tag{8.23}$$

Il modello autoregressivo specifica che la variabile di output dipende linearmente dai propri valori precedenti e da un termine stocastico; quindi il modello ha la forma di un'equazione alle differenze stocastiche (o relazione di ricorrenza). Insieme al modello a media mobile (MA), è un caso speciale e componente chiave del più generale autoregressivo-media mobile (ARMA) e modello autoregressivo integrato a media mobile modelli (ARIMA) di serie temporali. Tutti questi modelli sono applicabili solo nel caso in cui i dati siano temporalmente equi-distanziati.

Si può legittimamente infatti che la quantità di rifiuto raccolta in un sito dipende fortemente dal valore precedentemente registrato, quando era stato svuotato l'ultima volta il cassonetto. I dati a disposizione sono però, per loro natura, misurati a distanza di giorni variabile: applicare un modello autoregressivo poissoniano non è possibile. Per introdurre la forte componente temporale, è necessario definire un processo non discreto nel tempo ma bensì continuo: il processo di Ornstein-Uhlenbeck.

Il processo di Ornstein-Uhlenbeck [35] è un processo stazionario di Gauss-Markov: gaussiano, markoviano, temporalmente omogeneo. Il processo è nato come applicazione alla fisica, come modello per la velocità di una particella browniana massiccia sotto l'influenza dell'attrito. Nel tempo, il processo tende a spostarsi verso la sua funzione di media (mean-reverting) :

$$dY_t = -\phi Y_t + \sigma dW_t \tag{8.24}$$

dove $\phi > 0$ e W_t è il processo di Wiener.

Il processo può essere considerato una modifica del cammino casuale in tempo continuo in cui vi sia la tendenza del cammino a tornare indietro verso una posizione centrale, con un maggiore attrazione quando il processo è più lontano dal centro. Il processo di Ornstein-Uhlenbeck può anche essere considerato come l'analogo a tempo continuo del processo AR(1) a tempo discreto, considerando la distanza temporale non costante tra le osservazioni.

Il valore atteso è media ponderata (con pesi di somma 1) del valore iniziale e del livello di equilibrio di lungo periodo. Il peso del valore iniziale decresce esponenzialmente (tendendo a 0) al crescere di t.

Assumendo che δ_{ih} sia la distanza in giorni tra l'osservazione i e l'osservazione h, la correlazione segue la legge:

$$Corr(\theta_{ij}, \theta_{hj}) = \frac{{\sigma_b}^2}{\sigma_b^2 + \sigma_\epsilon^2} \exp(-\delta_{ih}\rho) \tag{8.25}$$

con $\rho > 0$.

Questo modello può essere riscritto come nell'equazione 8.21 con varianza totale $\sigma^2 = \sigma_b^2 + \sigma_\epsilon^2$ e matrice di correlazione come in 8.22, con correlazioni tra i ed h pari a $\phi_{i,h} = \frac{{\sigma_b}^2}{\sigma_b^2+\sigma_\epsilon^2} \exp(-\delta_{ih}\rho)$: l'unica differenza con il caso AR(1) è che il parametro è espresso in scala logaritmica, $\rho = -\log\phi$

8.5 Correlazione Spaziale

Nelle analisi statistiche in cui le misurazioni sono avvenute in punti diversi dello spazio, è bene includere anche un termine di correlazione spaziale: è infatti lecito supporre che rilevazioni avvenute in punti vicini nel territorio siano tra di loro simili, in quanto ad esempio facenti parte del medesimo complesso residenziale o situati nel medesimo quartiere. Come evidenziato in precedenza, all'interno di uno stesso comune potrebbero esserci sia zone industriali che residenziali, maggiormente o scarsamente abitate, che raccolgono diverse tipologie e quantità di rifiuti. Allo stesso tempo, però, ci sono comuni limitrofi come Rivoli e Collegno i cui siti appartengono anche alla medesima strada: è proprio in quest'ottica che la correlazione spaziale diventa essenziale nel modello predittivo. Decido di implementare solo correlazioni e covarianze stazionarie, che non dipendono da dove si trovino effettivamente i punti ma solo dalla reciproca distanza. Uso inoltre la distanza euclidea: questa è isotropica, ovvero non dipende dalla direzione.

Il valore di distanza d per cui non si ha più una correlazione tra le due variabili nello spazio viene denominato range, mentre il valore limite raggiunto, talvolta asintoticamente, viene denominato sill. [36]

Modello esponenziale La covarianza esponenziale raggiunge il valore del sill solo asintoticamente e di conseguenza non ammette range:

$$Cov(d) = \sigma^2 \exp\left\{\frac{-d}{\rho}\right\} \tag{8.26}$$

dove ρ è un parametro per scalare.

Modello gaussiano Presenta caratteristiche simili al modello precedente per quanto riguarda i parametri di sill e range, ma ha un comportamento presso l'origine che consente di rappresentare un fenomeno più regolare per distanze tra i siti prossime allo 0

$$Cov(d) = \sigma^2 \exp\left\{\frac{-d^2}{\rho}\right\} \tag{8.27}$$

Modello Matérn Questo modello risulta intermedio tra quello esponenziale (di cui è un caso particolare quando k = 0.5) e quello gaussiano e consente una certa flessibilità data dal parametro k:

$$Cov_k(d) = \sigma^2 \left\{\frac{2^{(1-k)}}{\Gamma(k)} \left(\frac{d}{\rho}\right)^k \mathcal{K}_k \left(\frac{d}{\rho}\right)\right\} \tag{8.28}$$

dove $\mathcal{K}$ è la funzione di Bessel del secondo tipo, che risolve l'equazione di Bessel per le armoniche cilindriche; Γ è la funzione gamma. Quando $k \to \infty$ la covarianza di Matérn converge alla funzione di covarianza esponenziale quadratica:

$$\lim_{k\to\infty} Cov_k(d) = \sigma^2 \exp\left\{\frac{-d^2}{2\rho^2}\right\} \tag{8.29}$$

8.6 Valutazione e Confronto Modelli

Esistono numerosi metodi per valutazione e selezionare il modello migliore, di seguito ne illustro alcuni.

Massima verosimiglianza I parametri dei GLM normalmente vengono stimati con il metodo della massima verosimiglianza (ML).

Sia $\mathcal{F}$ un modello statistico parametrico per i dati y. Gli elementi di $\mathcal{F}$ sono funzioni di densità di probabilità nel caso continuo o tutti funzioni di probabilità in quello discreto. In entrambi i casi si può scrivere

$$\mathcal{F} = \{p_y(y;\theta), \theta \in \Theta \subseteq \mathbb{R}^n\} \tag{8.30}$$

dove θ è un parametro n-dimensionale con valori nello spazio parametrico $\Theta \subseteq \mathbb{R}^n$. y assume valori nel supporto di Y sotto θ.

Si assume che θ sia identificabile, ossia che la corrispondenza tra Θ e $\mathcal{F}$ sia biunivoca. Sia $p^0(y)$ la vera e ignota densità di Y. Il modello $\mathcal{F}$ è detto correttamente specificato se $p^0(y) \subseteq \mathcal{F}$. Se $\mathcal{F}$ è correttamente specificato, il valore θ^0 tale che $p^0(y) = p_y(y;\theta)$ è detto vero valore del parametro. La funzione $\mathcal{L} : \Theta \to \mathbb{R}_+$ definita da

$$\mathcal{L}(\theta) = c(y)p_y(y;\theta) \tag{8.31}$$

con c(y) >0 costante non dipendente da θ, è detta funzione di verosimiglianza di θ basata sui dati y.

Spesso le procedure di inferenza basate su $\mathcal{L}$ sono espresse tramite la funzione di log-verosimiglianza:

$$l(\theta) = \log \mathcal{L}(\theta) \tag{8.32}$$

La funzione di verosimiglianza sintetizza l'informazione disponibile su θ alla luce dei dati y. Permette di confrontare l'adeguatezza, alla luce dei dati, di coppie di valori parametrici, θ' e θ'' in Θ, tramite il rapporto di verosimiglianza $\mathcal{L}(\theta')/\mathcal{L}(\theta'')$.

La funzione di verosimiglianza contiene tutta l'informazione su θ portata da y, fissato $\mathcal{F}$.

Un valore $\theta^* \subseteq \Theta$ tale che

$$\mathcal{L}(\theta*) \geq \mathcal{L}(\theta) \tag{8.33}$$

per ogni $\theta \subseteq \Theta$ è detto stima di massima verosimiglianza di θ. θ^* può essere determinato, di solito più agevolmente, anche utilizzando la funzione di log-verosimiglianza, essendo il logaritmo una funzione strettamente monotona crescente. Se $\omega(\theta)$ è una riparametrizzazione, essendo la funzione di verosimiglianza anche invariante, vale la proprietà di equivarianza dello stimatore di massima verosimiglianza: $\omega* = \omega(\theta^*)$

AIC Un approccio per la selezione del modello, diverso dal test del rapporto di verosimiglianza, è basato su penalizzazioni della log-verosimiglianza. L'AIC (Akaike Information Criterion) rappresenta un punteggio numerico che può essere utilizzato per determinare quale di più modelli è più probabile che sia il modello migliore per un dato set di dati. Stima i modelli relativamente, il che significa che i punteggi AIC sono utili solo rispetto ad altri punteggi AIC per lo stesso set di dati.

Tra due modelli, un punteggio AIC inferiore è preferibile. Esso viene calcolato dalla seguente formula:

$$AIC = -2\log l(\theta^*) + 2k \tag{8.34}$$

dove k il numero di parametri (o gradi di libertà).

BIC Altro metodo per la selezione del modello tra un insieme finito di modelli è il BIC (Bayesian Information Criterion).

Si basa, in parte, sulla funzione di verosimiglianza ed è strettamente correlato al criterio informativo di Akaike (AIC), e risolve possibili problemi di sovra-adattamento.

Il BIC introduce un termine di penalità per il numero di parametri nel modello, che risulta maggiore rispetto a quello utilizzato nell'AIC.

La formula di calcolo di questo stimatore è la seguente:

$$BIC = -2\log l(\theta^*) + nk\log n \tag{8.35}$$

dove n è il numero di dati.

Test anova L'analisi della varianza ANOVA (Analysis of Variance), è un insieme di tecniche statistiche inferenziali che permettono di confrontare due o più gruppi di dati confrontando la variabilità interna a questi gruppi con la variabilità tra i gruppi.

L'ipotesi nulla solitamente prevede che i dati di tutti i gruppi abbiano la stessa media, ovvero la stessa distribuzione stocastica, e che le differenze osservate tra i gruppi siano dovute solo al caso.

L'ipotesi alla base dell'analisi della varianza è che dati G gruppi, sia possibile scomporre la varianza in due componenti: Varianza interna ai gruppi (anche detta Varianza Within) e Varianza tra i gruppi (Varianza Between). La ragione che spinge a compiere tale distinzione è la convinzione che determinati fenomeni trovino spiegazione in caratteristiche proprie del gruppo di appartenenza. L'analisi della varianza si usa per determinare se più gruppi possono essere in qualche modo significativamente diversi tra loro (la varianza between contribuisce significativamente alla varianza totale - il fenomeno è legato a caratteristiche proprie di ciascun gruppo come il comune di appartenenza per il sito) o, viceversa, risultano omogenei (la varianza within contribuisce significativamente alla varianza totale - il fenomeno è legato a caratteristiche proprie di tutti i gruppi). In altre parole, il confronto si basa sull'idea che se la variabilità interna ai gruppi è relativamente elevata rispetto alla variabilità tra i gruppi, allora probabilmente la differenza tra questi gruppi è soltanto il risultato della variabilità interna.

Il più noto insieme di tecniche si basa sul confronto della varianza e usa variabili di test distribuite come la variabile casuale F di Fisher-Snedecor.

Questo test è utile quando si hanno modelli annidati per comprendere la significatività statistica dell'aggiungere o meno una variabile predittiva al modello.

Per verificare la significatività dell'intero modello si utilizza il test F. [37] Si abbia per semplicità un modello lineare completo:

$$Y = \beta_0 + \beta_1 X_1 + \dots \beta_k X_k + \epsilon \tag{8.36}$$

Y è un vettore di dimensione n e X, la matrice dei predittori, ha dimensione k (con $k < n$).

Sia SSE la somma delle differenze quadratiche tra i valori predetti dal modello lineare ($\hat{y}_i$) e i valori osservati y_i

$$SSE = \sum_{k=1}^{n} (\hat{y}_i - y_i)^2 \tag{8.37}$$

E sia SSR la somma delle differenze quadratiche tra i valori predetti e la media della variabile risposta $\bar{y}_i$

$$SSR = \sum_{k=1}^{n} (\hat{y}_i - \bar{y}_i)^2 \tag{8.38}$$

Si vuole verificare l'ipotesi nulla

$$H_0 := \beta_1 = \beta_2 = \dots = \beta_q = 0, \qquad q < k \tag{8.39}$$

Se fosse vera l'ipotesi nulla il modello sarebbe:

$$Y = \beta_0 + \beta_{q+1} X_{q+1} + \beta_{q+2} X_{q+2} + \dots \beta_k X_k + \epsilon \tag{8.40}$$

Siano SSR_r e SSE_r le somme quadratiche come sopra ma relative al modello ridotto. La statistica

$$\frac{(SSR - SSR_r)/q}{SSE/(n-k)} = \frac{(SSE_r - SSE)/q}{SSE/(n-k)} \tag{8.41}$$

sotto l'ipotesi nulla si distribuisce come una F con q ed n-k gradi di libertà.

Si rifiuta l'ipotesi nulla H_0 se $F > c$ in quanto, fissato α livello di significatività (di solito pari a 0.05), $P(F > c) = \alpha$.

Capitolo 9

Predizione dei Dati

Per il mio modello predittivo, integro il database delle misurazioni con altre informazioni di natura temporale e spaziale.

9.1 Componente Temporale

Immagino che la raccolta dei rifiuti sia influenzata dai seguenti parametri temporali:

- mese
- giorno della settimana e appartenenza o meno al weekend
- giorno festivo o feriale o pre/post festivo
- lockdown a causa della pandemia COVID-19

Infatti, è lecito pensare che durante il weekend venga raccolta più carta nei quartieri residenziali e che, al contrario, questo non avvenga per le zone industriali. Allo stesso modo, durante il periodo estivo molte aziende sono chiuse e i cittadini si spostano dalla propria abitazione. A causa del lockdown le abitudini dei cittadini sono cambiate: rimanendo in casa, i cittadini effettuavano molte delle proprie spese online, generando grandi quantità di rifiuti di imballaggi. Allo stesso modo però le attività commerciali ritenute non essenziali chiudevano.

Creo un'anagrafica dove, per ogni giorno dell'anno solare, salvo queste informazioni che ritengo possano influenzare la quantità di rifiuto prodotta dai cittadini. I cassonetti non vengono però svuotati giornalmente: è necessario che ogni giorno di raccolta raccolga al suo interno questi dati per tutti i giorni che sono intercorsi a partire dalla precedente rilevazione.

Integro così il db con i seguenti campi che ritengo possano agire da predittori per il modello:

- time2: giorno precedente in cui il cassonetto è stato svuotato
- n_gg: è l'esposto (o exposure) ovvero rappresenta il numero di giorni intercorsi tra la precedente ed attuale misurazione
- per ogni giorno della settimana *giorno_perc* calcolo la percentuale di n_gg che appartenevano a quel giorno della settimana
- covid_perc: calcolo la percentuale di giorni di lockdown sugli n_gg: ho scelto di individuare come periodo di lockdown i giorni dal 11 marzo 2020 al 18 maggio 2020 e dal 3 novembre 2020 fino al 30 aprile 2021
- agosto_perc: mostro, per il solo mese di agosto, la percentuale di giorni di quel mese sugli n_gg

- giorno: NumFactor per individuare il giorno e calcolato a partire dal 1 gennaio 2020
- group: colonna che ha sempre la valorizzazione 1. La inserisco perché per implementare le correlazioni spazio-temporali, è necessario definire un raggruppamento comune ai dati. In questo modo raggruppo insieme tutte le osservazioni

9.2 Componente Spaziale

Integro il database inserendo;

- giro1 che indica il macro-giro di cui fa parte un comune, che ricavo grazie all'utilizzo delle regole di associazione analizzate nei capitoli precedenti
- pos: coppia di coordinate geografiche (latitudine e longitudine) per calcolo delle distanze inter sito

Queste informazioni mi serviranno per decidere come raggruppare i dati e se calcolare le correlazioni spaziali solo all'interno di un comune, di un sotto o macro-giro. Inoltre implementerò tutti e 3 i metodi di correlazione spaziale presentati nel capitolo precedente.

A questo punto il database integrato, chiamato ***data_sett_sito***, ha questo header:

Listing 9.1. R output

```
##           id_sito       time   plate   rifiuto  comune_cod.y peso_tot      time2     n_gg
## 1 INTPAPPUB0023923 2021-03-30 FZ709LD  200101        000155       19     2021-03-27    3
## 2 INTPAPPUB0023923 2020-12-31 FZ338SC  200101        000155        0     2020-12-29    2
## 3 INTPAPPUB0023923 2020-12-15 FN675FH  200101        000155       43     2020-12-12    3
## 4 INTPAPPUB0023923 2021-02-09 FN675FH  200101        000155       34     2021-02-06    3
## 5 INTPAPPUB0023923 2021-02-02 FN675FH  200101        000155       26     2021-01-30    3
## 6 INTPAPPUB0023923 2021-02-27 EL850BV  200101        000155       14     2021-02-09   18

##              pos       capacita      trans     attivo           giro1
## 1 (45.07184,7.55772)     5000     2020-11-26 2020-11-26 ColVenDruSanPia
## 2 (45.07184,7.55772)     5000     2020-11-26 2020-11-26 ColVenDruSanPia
## 3 (45.07184,7.55772)     5000     2020-11-26 2020-11-26 ColVenDruSanPia
## 4 (45.07184,7.55772)     5000     2020-11-26 2020-11-26 ColVenDruSanPia
## 5 (45.07184,7.55772)     5000     2020-11-26 2020-11-26 ColVenDruSanPia
## 6 (45.07184,7.55772)     5000     2020-11-26 2020-11-26 ColVenDruSanPia

##   mar_perc  mer_perc  gio_perc  ven_perc agosto_perc covid_perc fest_perc
## 1 0.3333333 0.0000000 0.0000000 0.0000000           0          1         0
## 2 0.0000000 0.5000000 0.5000000 0.0000000           0          1         0
## 3 0.3333333 0.0000000 0.0000000 0.0000000           0          1         0
## 4 0.3333333 0.0000000 0.0000000 0.0000000           0          1         0
## 5 0.3333333 0.0000000 0.0000000 0.0000000           0          1         0
## 6 0.1111111 0.1666667 0.1666667 0.1666667           0          1         0

##     giro2 group giorno  dom_perc  lun_perc
## 1 Collegno   1  (456)  0.3333333 0.3333333
## 2 Collegno   1  (367)  0.0000000 0.0000000
## 3 Collegno   1  (351)  0.3333333 0.3333333
## 4 Collegno   1  (407)  0.3333333 0.3333333
## 5 Collegno   1  (400)  0.3333333 0.3333333
## 6 Collegno   1  (425)  0.1111111 0.1111111
```

Fitto vari modelli con effetti autoregressivi [38] e li confronto utilizzando l'indicatore AIC, BIC e log-verosimiglianza.

Fitto il modello considerando usando:

- componente temporale con modello autoregressivo continuo tramite comando *ou*
- comune_cod.y è un factor usato come variabile regressiva
- tot_avg come variabile regressiva

- *giorno_perc* come variabili regressive. Non inserisco il regressore sab_perc perché ovviamente la percentuale degli n_gg passati che erano sabato si ottiene di differenza a partire dalle altre percentuali, togliendo le valorizzazioni per tutti gli altri giorni della settimana
- effetto del regressore agosto_perc
- effetto del regressore covid_perc
- autocorrelazione spaziale tramite posizione e distanza tra i siti
- $\log(n\_gg)$ come offset
- funzione link logaritmica
- regressione di Poisson

9.3 Stima GLMM in R

Esistono diversi pacchetti in R per la stima dei GLMM: determinare quale metodo di stima non è un'impresa semplice. Mostro qui un elenco dei principali metodi e pacchetti che permettono di implementare l'algoritmo con diversi stimatori numerici [39]:

Metodo	Vantaggi	Svantaggi	Pacchetto
quasi-verosimiglianza penalizzata	flessibile, largamente implementata	bias per grandi varianze o piccole medie	MASS(glmmPQL)
approssimazione Laplace	più accurata di PQL	più lenta e meno flessibile di PQL	lme4(glmer), GlmmADMB, INLA, glmmmTMB
Formula di quadratura di Gauss-Hermite	più accurata di Laplace	più lenta di Laplace, limitata tra 2-3 effetti casuali	lme4(glmer), glmmML
MCMC (Catena di Markov Monte Carlo)	altamente flessibile, accurata, numero arbitrario di effetti casuali	molto lenta	MCMCglmm, rstanatm, brms, MCMCpack, JAGS, glmmADMB

Tabella 9.1. Elenco Metodi di Stima Parametri GLMM

Ho deciso per questo elaborato di utilizzare il pacchetto R *glmmTMB* [24] poiché, nonostante il metodo di approssimazione di Laplace sia lento, permette di inserire correlazioni temporali a tempo continuo e correlazioni spaziali. Creo 9 modelli, che analizzerò tramite i criteri per la model selection discussi nel capitolo precedente.

Modello 1 In questo modello non inserisco le variabili agosto_perc e covid_perc. La covarianza spaziale segue il modello esponenziale e non c'è alcun ulteriore raggruppamento: tutti i dati condividono gli stessi parametri per la covarianza.

La componente di autocorrelazione temporale segue il processo continuo di Ornstein-Uhlenbeck *ou* [40] i cui parametri per la covarianza sono gli stessi per tutto il dataset.

Listing 9.2. modello 1

```
library(glmmTMB)

f1 <- glmmTMB(peso_tot ~ comune_cod.y
                         + ou (giorno + 0| group )
                         + tot_tag
                         + dom_perc + lun_perc + mar_perc + mer_perc + gio_perc + ven_perc
                         + exp(pos+ 0 | group)
                         + offset(log(n_gg  )),
                         data=data_sett_sito,
                         family = poisson(link = "log") )
```

Modello 2 In questo modello non inserisco le variabili agosto_perc e covid_perc. La covarianza spaziale segue il modello esponenziale e non c'è alcun ulteriore raggruppamento. La componente ou raggruppa per comune: all'interno dello stesso comune, tutti i parametri della covarianza temporale saranno i medesimi.

Listing 9.3. modello 2

```
f2 <- glmmTMB(peso_tot ~ comune_cod.y
                         + ou (giorno + 0| comune_cod.y )
                         + tot_tag
                         + dom_perc + lun_perc + mar_perc + mer_perc + gio_perc + ven_perc
                         + exp(pos+ 0 | group)
                         + offset(log(n_gg  )),
                         data=data_sett_sito,
                         family = poisson(link = "log") )
```

Modello 3 In questo modello non inserisco le variabili agosto_perc e covid_perc. La covarianza spaziale segue il modello esponenziale e non raggruppa. La componente temporale raggruppa per sito.

Listing 9.4. modello 3

```
f3 <- glmmTMB(peso_tot ~ comune_cod.y
                         + ou (giorno + 0| id_sito )
                         + tot_tag
                         + dom_perc + lun_perc + mar_perc + mer_perc + gio_perc + ven_perc
                         + exp(pos+ 0 | group)
                         + offset(log(n_gg  )),
                         data=data_sett_sito,
                         family = poisson(link = "log") )
```

Modello 4 In questo modello non inserisco le variabili agosto_perc e covid_perc. La covarianza spaziale segue il modello esponenziale e raggruppa per comune. La componente ou raggruppa per sito.

Listing 9.5. modello 4

```
f4 <- glmmTMB(peso_tot ~ comune_cod.y
                         + ou (giorno + 0| id_sito )
                         + tot_tag
                         + dom_perc + lun_perc + mar_perc + mer_perc + gio_perc + ven_perc
                         + exp(pos + 0 | comune_cod.y)
                         + offset(log(n_gg  )),
                         data=data_sett_sito,
                         family = poisson(link = "log") )
```

Modello 5 In questo modello inserisco la variabile covid_perc. La covarianza spaziale segue il modello esponenziale e raggruppa per comune. La componente ou raggruppa per sito.

Listing 9.6. modello 5

```
f5 <- glmmTMB(peso_tot ~ comune_cod.y
                          + ou (giorno + 0| id_sito )
                          + tot_tag
                          + dom_perc + lun_perc + mar_perc + mer_perc + gio_perc + ven_perc
                          + covid_perc
                          + exp(pos + 0 | comune_cod.y)
                          + offset(log(n_gg  )),
                          data=data_sett_sito,
                          family = poisson(link = "log") )
```

Modello 6 In questo modello inserisco le variabili agosto_perc e covid_perc. La covarianza spaziale segue il modello esponenziale e raggruppa per comune. La componente ou raggruppa per sito.

Listing 9.7. modello 6

```
f6 <- glmmTMB(peso_tot ~ comune_cod.y
                          + ou (giorno + 0| id_sito )
                          + tot_tag
                          + dom_perc + lun_perc + mar_perc + mer_perc + gio_perc + ven_perc
                          + agosto_perc
                          + covid_perc
                          + exp(pos + 0 | comune_cod.y)
                          + offset(log(n_gg  )),
                          data=data_sett_sito,
                          family = poisson(link = "log") )
```

Modello 7 In questo modello non inserisco le variabili agosto_perc e covid_perc. La covarianza spaziale segue il modello esponenziale e raggruppa per giro1 (la macro suddivisione dei comuni in 2 cluster) . La componente ou raggruppa per sito. L'intercetta del modello lineare generalizzato dipende dal comune, che è il gruppo.

Listing 9.8. modello 7

```
f7 <- glmmTMB(peso_tot ~  (1|comune_cod.y)
                          + ou (giorno + 0| id_sito )
                          + tot_tag
                          + dom_perc + lun_perc + mar_perc + mer_perc + gio_perc + ven_perc
                          + exp(pos + 0 | giro1)
                          + offset(log(n_gg  )),
                          data=data_sett_sito,
                          family = poisson(link = "log") )
```

Modello 8 In questo modello non inserisco le variabili agosto_perc e covid_perc. La covarianza spaziale segue il modello di Matérn e non raggruppa. La componente ou raggruppa per sito.

Listing 9.9. modello 8

```
f3_mat <- glmmTMB(peso_tot ~ comune_cod.y
                             + ou (giorno + 0| id_sito )
                             + tot_tag
                             + dom_perc + lun_perc + mar_perc + mer_perc + gio_perc + ven_perc
                             + mat(pos+ 0 | group)
                             + offset(log(n_gg  )),
                             data=data_sett_sito,
                             family = poisson(link = "log"),
                             dispformula=~0 )
```

Modello 9 In questo modello non inserisco le variabili agosto_perc e covid_perc. La covarianza spaziale segue il modello di Matérn e raggruppa per comune. La componente ou raggruppa per sito.

Listing 9.10. modello 9

```
f4_mat<- glmmTMB(peso_tot ~ comune_cod.y
                          + ou (giorno + 0| id_sito )
                          + tot_tag
                          + dom_perc + lun_perc + mar_perc + mer_perc + gio_perc + ven_perc
                          + mat(pos+ 0 | comune_cod.y)
                          + offset(log(n_gg  )),
                          data=data_sett_sito,
                          family = poisson(link = "log"),
                          dispformula=~0 )
```

A questo punto per la scelta del modello migliore, mostro i valori confrontati di AIC ed ottengo questo output:

Listing 9.11. R output

```
Model selection based on AICc:

        K     AICc Delta_AICc AICcWt Cum.Wt        LL
f6     23 221316.8       0.00      1      1 -110635.4
f5     22 221487.0     170.16      0      1 -110721.5
f4_mat 22 221487.8     170.93      0      1 -110721.9
f3_mat 22 221488.5     171.63      0      1 -110722.2
f3     21 221499.1     182.29      0      1 -110728.5
f4     21 221499.1     182.29      0      1 -110728.5
f7     13 221512.5     195.65      0      1 -110743.2
f2     21 529777.2  308460.36      0      1 -264867.6
f1     21 616929.6  395612.77      0      1 -308443.8
```

Mentre con il BIC ho:

Listing 9.12. R output

```
Model selection based on BIC:

        K      BIC Delta_BIC BICWt Cum.Wt        LL
f6     23 221501.1      0.00     1      1 -110635.4
f7     13 221616.7    115.53     0      1 -110743.2
f5     22 221663.3    162.15     0      1 -110721.5
f4_mat 22 221664.0    162.92     0      1 -110721.9
f3_mat 22 221664.8    163.62     0      1 -110722.2
f3     21 221667.4    166.27     0      1 -110728.5
f4     21 221667.4    166.27     0      1 -110728.5
f2     21 529945.5 308444.34     0      1 -264867.6
f1     21 617097.9 395596.75     0      1 -308443.8
```

.

In entrambi i casi i modelli sono ordinati da quello preferibile (con indicatore più basso) a quello peggiore:

- K rappresenta il numero di parametri stimati per ogni modello (ovvero i gradi di libertà)
- AIC/BIC è l'indicatore di valutazione del modello
- Delta_AIC/Delta_BIC è la differenza di indicatore rispetto al modello da selezionare, quello con valori inferiori
- AICWt/BICWt sono i pesi, ovvero la probabilità/il livello di evidenza che il modello considerato sia quello migliore, da selezionare in quanto fitta meglio i dati presenti
- Cum.Wt è la somma cumulata di AICWt/BICWt

- LL è la log-verosimiglianza

Si nota che il modello migliore, f6, ha sia il valore inferiore di AIC che di BIC. Il modello ad effetti misti f7 ha un AIC relativamente alto ma BIC basso rispetto agli altri modelli: questo perché, pur avendo meno gradi di libertà, interpola bene i dati.

Questi criteri sembrano suggerirci che effettivamente il modello f6 sia quello da selezionare. Poiché in realtà i modelli f4, f5 e f6 sono annidati (tramite, rispettivamente, aggiunta delle variabili regressive covid_perc e, successivamente, anche agosto_perc), per verificare la significatività dei modelli effettuo il test di Fisher tramite anova a 2 vie.

Utilizzo la funzione *anovaOD* [44] presente nel pacchetto *AICcmodavg* [26]: questa funzione applica un correttivo al test di rapporto di verosimiglianza. Il test impone

$$LR = -2(LL_g - LL_p) \tag{9.1}$$

dove LL_g è la log-verosimiglianza del modello con più variabili e LL_p del modello più piccolo, con meno variabili predittive. In questo caso il test statistico si distribuisce come una variabile $\chi^2(K_g - K_p)$ in cui K_g sono i gradi di libertà / numero di osservazioni e nodi usati per l'analisi del modello grande e K_p il corrispettivo per il modello piccolo.

Confronto inizialmente f4 e f5:

Listing 9.13. R output

```
Data: data_sett_sito
Models:
f4: peso_tot ~ (comune_cod.y) + ou(giorno + 0 | id_sito) + tot_tag + , zi=~0, disp=~1
f4:     dom_perc + lun_perc + mar_perc + mer_perc + gio_perc + ven_perc + , zi=~0, disp=~1
f4:     exp(pos + 0 | comune_cod.y) + offset(log(n_gg)), zi=~0, disp=~1
f5: peso_tot ~ comune_cod.y + ou(giorno + 0 | id_sito) + tot_tag + , zi=~0, disp=~1
f5:     dom_perc + lun_perc + mar_perc + mer_perc + gio_perc + ven_perc + , zi=~0, disp=~1
f5:     covid_perc + exp(pos + 0 | comune_cod.y) + offset(log(n_gg)), zi=~0, disp=~1
   Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
f4 21 221499 221667 -110729   221457
f5 22 221487 221663 -110721   221443 14.134      1  0.0001702 ***
```

Poiché il p-value $Pr(> Chisq)$ è inferiore al livello di significatività $\alpha = 0.05$, allora rigetto l'ipotesi nulla in quanto c'è differenza statistica rilevante tra i due modelli.

Confronto ora f5 con f6:

Listing 9.14. R output

```
Data: data_sett_sito
Models:
f5: peso_tot ~ comune_cod.y + ou(giorno + 0 | id_sito) + tot_tag + , zi=~0, disp=~1
f5:     dom_perc + lun_perc + mar_perc + mer_perc + gio_perc + ven_perc + , zi=~0, disp=~1
f5:     covid_perc + exp(pos + 0 | comune_cod.y) + offset(log(n_gg)), zi=~0, disp=~1
f6: peso_tot ~ comune_cod.y + ou(giorno + 0 | id_sito) + tot_tag + , zi=~0, disp=~1
f6:     dom_perc + lun_perc + mar_perc + mer_perc + gio_perc + ven_perc + , zi=~0, disp=~1
f6:     covid_perc + agosto_perc + exp(pos + 0 | comune_cod.y) + , zi=~0, disp=~1
f6:     offset(log(n_gg)), zi=~0, disp=~1
   Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
f5 22 221487 221663 -110721   221443
f6 23 221317 221501 -110635   221271 172.17      1  < 2.2e-16 ***
```

Anche in questo caso rifiuto l'ipotesi nulla e ritengo statisticamente significativa l'aggiunta delle variabili agosto_perc e covid_perc al modello predittivo.

Per vedere se i regressori inseriti in tutti i modelli predittivi sono effettivamente significativi, mostro parte del *summary* del modello f6:

Listing 9.15. R output

```
Family: poisson  ( log )
Formula:
peso_tot ~ comune_cod.y + ou(giorno + 0 | id_sito) + tot_tag +
    dom_perc + lun_perc + mar_perc + mer_perc + gio_perc + ven_perc +
    covid_perc + agosto_perc + exp(pos + 0 | comune_cod.y) +
    offset(log(n_gg))
Data: data_sett_sito

     AIC       BIC    logLik  deviance  df.resid
 221316.8  221501.1 -110635.4  221270.8     22343

Number of obs: 22366, groups:  id_sito, 316; comune_cod.y, 10

Conditional model:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         4.342e-01  2.556e-01   1.699 0.089344 .
comune_cod.y000153  7.223e-01  2.459e-01   2.938 0.003308 **
comune_cod.y000154  9.143e-01  2.302e-01   3.973 7.11e-05 ***
comune_cod.y000155  6.450e-01  2.306e-01   2.797 0.005163 **
comune_cod.y000157  6.410e-01  2.810e-01   2.281 0.022542 *
comune_cod.y000158  1.633e+00  4.260e-01   3.832 0.000127 ***
comune_cod.y000161  8.040e-01  4.760e-01   1.689 0.091235 .
comune_cod.y000162  1.030e+00  2.525e-01   4.078 4.54e-05 ***
comune_cod.y000277  1.062e+00  2.439e-01   4.353 1.34e-05 ***
comune_cod.y001150  6.107e-01  2.515e-01   2.428 0.015199 *
tot_tag             5.498e-01  1.121e-01   4.904 9.38e-07 ***
dom_perc            1.834e-01  1.956e+03   0.000 0.999925
lun_perc            1.834e-01  1.956e+03   0.000 0.999925
mar_perc           -8.721e-02  7.922e-02  -1.101 0.270942
mer_perc            1.056e+00  9.562e-02  11.039  < 2e-16 ***
gio_perc           -9.430e-01  1.115e-01  -8.457  < 2e-16 ***
ven_perc            2.078e-02  1.193e-01   0.174 0.861717
covid_perc          3.352e-03  1.341e-02   0.250 0.802650
agosto_perc        -3.715e-01  2.813e-02 -13.208  < 2e-16 ***
```

Come si può notare, non sono significative le variabili con p_value molto alti, bel al di sopra del livello di significatività $\alpha = 0.05$: le elimino allora dal modello, che diventa:

Listing 9.16. modello 6 aggiornato

```
f6 <- glmmTMB(peso_tot ~ comune_cod.y
                          + ou (giorno + 0| id_sito )
                          + tot_tag
                          + mer_perc + gio_perc
                          + agosto_perc
                          + exp(pos + 0 | comune_cod.y)
                          + offset(log(n_gg  )),
                          data=data_sett_sito,
                          family = poisson(link = "log") )
```

il cui *summary* è:

Pur avendo eliminato 5 variabili predittive (dom_perc, lun_perc, mar_perc, ven_perc e covid_perc), l'AIC è aumentato di pochissimo, passando da 221316.8 a 221317.5. La variabile inoltre che portava veramente incremento al momento non era il correttivo del periodo di quarantena, ma l'individuazione del mese di ferie di agosto. Adesso l'unica variabile che non è ritenuta significativa è la variabile *dummy* che individua l'effetto di appartenenza al comune_cod "000161", ovvero Alpignano: è proprio questo l'unico comune che ha fatto parte, per una frazione di giorni, di entrambi i macro-giri di raccolta e che, come mostrato in figura 7.6 "collega" virtualmente i due macro-cluster dei siti.

Listing 9.17. R output

```
Family: poisson  ( log )
Formula:
peso_tot ~ comune_cod.y + ou(giorno + 0 | id_sito) + tot_tag +
    mer_perc + gio_perc + agosto_perc + exp(pos + 0 | comune_cod.y) +
    offset(log(n_gg))
Data: data_sett_sito

      AIC       BIC    logLik  deviance  df.resid
 221317.5  221461.8 -110640.8  221281.5     22348

Number of obs: 22366, groups:  id_sito, 316; comune_cod.y, 10

Conditional model:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        0.52223    0.24626   2.121 0.033950 *
comune_cod.y000153  0.72161    0.24581   2.936 0.003328 **
comune_cod.y000154  0.91458    0.23011   3.975 7.05e-05 ***
comune_cod.y000155  0.64590    0.23058   2.801 0.005092 **
comune_cod.y000157  0.64037    0.28098   2.279 0.022661 *
comune_cod.y000158  1.63563    0.42598   3.840 0.000123 ***
comune_cod.y000161  0.80431    0.47601   1.690 0.091088 .
comune_cod.y000162  1.02805    0.25247   4.072 4.66e-05 ***
comune_cod.y000277  1.05980    0.24380   4.347 1.38e-05 ***
comune_cod.y001150  0.61022    0.25152   2.426 0.015260 *
tot_tag             0.54973    0.11210   4.904 9.39e-07 ***
mer_perc            0.95746    0.06675  14.344  < 2e-16 ***
gio_perc           -1.14272    0.07277 -15.703  < 2e-16 ***
agosto_perc        -0.37309    0.02715 -13.744  < 2e-16 ***
```

Capitolo 10

Conclusioni

10.1 Considerazioni Finali

Trovato l'algoritmo predittivo ottimale tra quelli proposti, posso utilizzare questo modello per effettuare previsioni sul comportamento futuro delle quantità raccolte nei cassoneti, avendo cura di pre-processare i dati ed integrarli come mostrato nei capitoli precedenti.

Come evidenziato da questo elaborato di tesi, trattare dati reali è molto complesso: la conoscenza del dominio è lo strumento essenziale per permettere di integrare coerentemente le informazioni a disposizione. I dati reali sono tra loro eterogenei e rispecchiano la forte componente caotica della realtà.

10.2 Possibili Implementazioni Future

Sono molteplici gli approfondimenti applicabili all'analisi:

- come prima estensione, si potrebbero considerare non solo i rifiuti cartacei ma anche delle altre tipologie di scarto. Il tipo di rifiuto diventerebbe una ulteriore variabile predittiva nel modello. In questo modo si introdurrebbe anche una correlazione tra le diverse tipologie di scarto, che potrebbe permette di differenziare e clusterizzare meglio i siti dalle caratteristiche similari

- si potrebbe effettuare un'analisi sulla zona di locazione del sito, integrando a monte l'informazione se si trovi in un quartiere residenziale oppure in un complesso industriale per utilizzare questo dato come regressore

- al posto dei modelli d'inferenza statistica, si potrebbe studiare un modello gerarchico di tipo bayesiano. Nello specifico, il pacchetto INLA (Integrated Nested Laplace Approximation) è particolarmente utile nell'integrazione di correlazioni spazio-temporali

- l'algoritmo potrebbe girare in real-life: integrando le misurazioni effettuate dal veicolo ed inserendole subito nel database (a seguito di operazioni di pre-processing e pulizia dei dati automatizzate): in questo modo, se ad esempio le quantità raccolte nei primi m dei totali n siti da raggiungere in giornata fosse inferiore allo stimato, si potrebbe modificare in tempo reale il giro di raccolta del veicolo. Al contrario, se la quantità raccolta fosse superiore allo stimato per quei m siti, allora l'atteso dei siti vicini e correlati potrebbe superare la soglia minima richiesta per lo svuotamento e quei cassonetti potrebbero essere introdotti nel giro di raccolta.

- in questo estratto, non ho considerare le misurazioni pari a 0 come outlier o valori mancanti: questo perché erano presenti molte valorizzazioni a peso nullo, e considerate come dovute al cassonetto effettivamente vuoto: un'implementazione successiva potrebbe trattare questi

dati come valori mancanti, segnalando che la misurazione della quantità non è andata a buon fine ma che il cassonetto è effettivamente stato svuotato

- l'analisi potrebbe permettere di individuare in quali aree geografiche posizionare nuovi cassonetti, definendo quali sono le zone del territorio dove si raccolgono le maggiori quantità di rifiuti e clusterizzando i siti usando algoritmi basati sul peso totale come variabile di densità
- al posto del peso, si potrebbe modellizzare il problema utilizzando come variabile risposta il volume totale raccolto. Per fare questo, però, è necessario integrare i dati con le stime della densità dei rifiuti e dedurre o aggiungere informazioni sulla natura del territorio e delle proprie aree residenziali o industriali

Bibliografia

[1] https://www.avvenire.it/mondo/pagine/clima-2016-record-caldo

[2] https://www.rinnovabili.it/.../co2-in-atmosfera-record-2021/

[3] https://unric.org/it/agenda-2030/

[4] https://www.falacosagiusta.org/2021/economia-circolare-litalia-e-tra-i-leader-in-europa/

[5] https://cidiu.it/wp-content/uploads/2021/08/BdS2020_web_ok.pdf

[6] https://www.fasda.it/economia-circolare-rifiuti/

[7] https://www.europarl.europa.eu/news/it/headlines/priorities/economiacircolare

[8] https://www.nonsoloambiente.it/economia-circolare-cradle-to-cradle

[9] https://www.abenergie.it/blog/2021/05/la-raccolta-differenziata-in-italia-e-lo-sviluppo-sostenibile-siamo-uneconomia-davvero-circolare

[10] https://www.catasto-rifiuti.isprambiente.it

[11] https://www.welfarenetwork.it/

[12] https://www.comieco.org/

[13] https://theprospectornews.com/

[14] https://cidiu.it/

[15] https://www.nordengineering.com/

[16] https://www.fincantierinxt.it/index.php/it/azienda-ita

[17] https://onde.city/

[18] https://moltosenso.com/ottimizzazione-della-raccolta-differenziata/

[19] James G., Witten D., Hastie T., Tibshirani R., *An Introduction to Statistical Learning with Applications in R*, ISBN 13:9781461471370

[20] https://www.rdocumentation.org/packages/sqldf/versions/0.4-11

[21] https://www.rdocumentation.org/packages/ggmap/versions/3.0.0

[22] https://cran.r-project.org/web/packages/geosphere/index.html

[23] https://www.rdocumentation.org/packages/arules/versions/1.7-0

[24] https://cran.r-project.org/web/packages/glmmTMB/index.html

[25] https://cran.r-project.org/web/packages/glmmTMB/vignettes/covstruct.html

[26] https://cran.r-project.org/web/packages/AICcmodavg/index.html

[27] https://www.rdocumentation.org/packages/geosphere/versions/1.5-14/topics/distHaversine

[28] http://dbdmg.polito.it/wordpress/wp-content/uploads/2010/12/14-DMregole_associazione6.pdf

[29] https://cran.r-project.org/doc/contrib/Ricci-regression-it.pdf

[30] Salvan A, Sartori N., Pace L (2020), *Modelli Lineari Generalizzati*

[31] https://www.unica.it/static/resources/cms/documents/Regressione_Poisson.pdf

[32] Zuur, A., Ieno, E. N., Walker, N., Saveiliev, A. A., and Smith, G. M. (2009). *Mixed Effects Models and Extensions in Ecology with R.* Springer, New York. ISBN 978-0-387-87457-9.

[33] https://online.stat.psu.edu/stat462/node/189/

[34] https://arxiv.org/pdf/1902.08828.pdf

[35] https://investmentmath.com/math/2013/12/01/from-gaussian-to-ornstein-uhlenbeck-processes.html

[36] http://tesi.cab.unipd.it/10076/1/Levorato_Daniele.pdf

[37] http://www00.unibg.it/dati/corsi/8910/24007-lez_08.pdf

[38] https://becarioprecario.bitbucket.io/inla-gitbook/ch-temporal.html

[39] https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html#what-methods-are-available-to-fit-estimate-glmms

[40] https://inla.r-inla-download.org/r-inla.org/doc/latent/ou.pdf

[41] Pace, L., Salvan, A.(2001). *Introduzione alla statsitica. II Inferenza, verosimiglianza, modelli.* CEDAM, Padova.

[42] S.N. Wood, *Generalized additive models an introduction with R*, Chapman and Hall, 2006

[43] H. Wakernagel, *Multivariate geostatistics : An introduction with applications.*, Springer, Berlin, 2003.

[44] https://search.r-project.org/CRAN/refmans/AICcmodavg/html/anovaOD.html