POLITECNICO DI TORINO

Corso di Laurea Magistrale in Ingegneria Civile

Tesi di Laurea Magistrale

Previsione della congestione stradale in ambito urbano e extra-urbano: un'applicazione dei big data e del machine learning

Relatore:

Prof. Deflorio Francesco

Candidato:

Sibille Alberto - 242174

Anno Accademico 2018/2019

Ringraziamenti

Questa tesi è stata sviluppata durante uno stage di 6 mesi presso gli uffici di Setec International a Parigi, nell'ambito di un progetto Erasmus che coinvolge il Politecnico di Torino e l'École Nationale des Ponts et Chaussées di Parigi. Tale lavoro è il risultato di una collaborazione con docenti (italiani e francesi) e professionisti che hanno dedicato tempo ed energie a questo progetto. Per tale ragione desidero ringraziarli personalmente.
Ringrazio il Prof. Francesco Deflorio, relatore di questa tesi presso il Politecnico di Torino, per la sua grande disponibilità nel seguirmi a distanza durante la realizzazione e la redazione della tesi e per i preziosi consigli dispensati.
Ringrazio inoltre il Prof. François Combes, relatore di questa tesi presso l'ENPC di Parigi, per avermi accompagnato nella fasi iniziali della sperimentazione.
Infine un sentito ringraziamento a Didier Revillon, che mi ha offerto l'opportunità di far parte della sua equipe e che crede fortemente in questo progetto, ad Aric Wizenberg, che con grande pazienza mi ha iniziato al mondo della data science, e a tutto il dipartimento "Mobilità e Trasporti" di Setec International.

Ringrazio i miei genitori per essere sempre stati presenti durante il mio percorso accademico, consigliandomi al meglio nelle scelte difficili ma soprattutto appoggiandomi in qualsiasi decisione infine presa. Grazie anche per avermi dimostrato che qualsiasi sfida, se affrontata insieme, si può superare.

Ringrazio Giulia per essere stata "al mio fianco" in ogni momento, anche e soprattutto a 600 chilometri di distanza, e per avermi sempre sostenuto con pazienza e affetto. Grazie perché nei momenti difficili so a chi rivolgermi per ritrovare il sorriso.

Grazie a tutti i miei amici, a quelli di lunga data e a quelli più recenti, con cui ho condiviso dei momenti felici in questi cinque anni; c'è un pezzo di tutti voi in questa tesi.

Abstract

Questo progetto di tesi ha come obiettivo lo sviluppo di uno strumento per la previsione della congestione stradale in tempo reale ad un orizzonte temporale di 30 minuti. La previsione della congestione viene affrontata attraverso un approccio *data-driven*. L'interesse di questo approccio risiede nella possibilità di applicare all'ingegneria del traffico gli algoritmi di *machine learning*: tali algoritmi permettono di sfruttare la mole sempre crescente di dati di mobilità di cui disponiamo oggi, i *big data*, senza la necessità di costruire modelli matematici che tentano di riprodurre gli aspetti complessi di comportamento del sistema. L'obiettivo è la creazione di uno strumento applicabile sia in ambito urbano che in ambito extra-urbano.

Nella prima parte dello studio vengono analizzati e confrontati i dati potenzialmente utilizzabili per alimentare il modello: misure di portata e tasso di occupazione rilevati tramite spire induttive e misure di velocità media rilevate tramite Floating Car Data (FCD).

Durante l'analisi, le serie osservate sono disaggregate in una componente media, basata su dati storici e che quindi non necessita di essere predetta, ed una componente variabile, che sarà l'oggetto della previsione. Il fenomeno è descritto mediante alcune variabili esplicative (temporali, spaziali e metereologiche) che sono fornite al modello per essere elaborate.

Attraverso l'applicazione di algoritmi di apprendimento supervisionato si ottengono previsioni dei valori di flusso di traffico e tasso di occupazione ad un orizzonte temporale di 30 minuti, per alcuni segmenti stradali rappresentativi; tali valori sono confrontati con le misure osservate per valutare le prestazioni del modello.

Infine, a partire dai valori predetti, si cerca di determinare se tali condizioni di traffico comportino o meno l'insorgere di fenomeni locali di congestione stradale.

In questo progetto, il modello è testato su due reti stradali di tipologia differente: quella urbana della città di Lione e quella nazionale extra-urbana dell'Ile-de-France.

Abstract

The purpose of this work is the development of a tool for predicting road congestion in real time at a time horizon of 30 minutes. Congestion forecasting is made through a data-driven approach. The interest of this approach lies in the possibility of applying machine learning algorithms to traffic engineering: these algorithms allow us to exploit the increasing amount of mobility data available today, the big data, without the need to build mathematical models that try to reproduce the complex behaviour of the system. The objective is the creation of a tool that can perform both in urban and extra-urban environments.

In the first part of the study, the potential input data are analysed and compared: flow and occupancy rate measurements taken by inductive loops and average speed measurements taken by Floating Car Data (FCD).

Subsequently, the observed series are decomposed into an average component based on historical data, which does not need to be predicted, and a variable component that will be the object of the forecast. The explanatory variables (temporal, spatial and meteorological) that will be provided to the model are then created.

Through the application of supervised learning algorithms, the values of range and occupancy rate, at a time horizon of 30 minutes for representative road segments, are forecasted; these values are compared with the observed values to evaluate the model's performance.

Finally, based on these values, an attempt is made to determine whether such traffic conditions lead to the emergence of road congestion.

In this project, the model is tested on two different road networks: the urban network of the city of Lyon and the national suburban network of Ile-de-France.

Indice

Indice delle figure

Indice delle tabelle

I. Introduzione

La previsione della congestione stradale è un argomento di grande interesse per l'ingegneria dei trasporti. La formazione di ingorghi sugli assi di accesso o uscita delle grandi città è un problema sempre attuale, che si verifica frequentemente in molte parti del mondo. Secondo uno studio condotto da INRIX, una società americana che fornisce servizi e statistiche legati alla circolazione stradale, le tre città più congestionate d'Europa sono Mosca, Londra e Parigi. [1]

Il tempo perso dagli automobilisti bloccati nel traffico rappresenta un costo per la collettività. In economia dei trasporti viene attribuito un valore al tempo degli automobilisti; ciò permette di valorizzare un'infrastruttura ben dimensionata rispetto ad una sottodimensionata. La perdita di tempo degli automobilisti nel traffico si trasforma quindi in una perdita di denaro per la collettività; al contrario una riduzione della congestione stradale ha come conseguenza un ritorno economico.

Non deve quindi stupire che le società concessionarie di infrastrutture stradali siano interessate ad investire in strumenti che permettano di prevedere, e quindi anticipare, la congestione al fine di offrire agli utenti un servizio migliore. Essere in grado di prevedere la formazione di un ingorgo stradale, con 30 minuti d'anticipo e con una buona precisione circa la sua localizzazione e durata, permetterebbe a questi operatori di mettere in atto misure per fronteggiare la congestione. Tali misure possono andare dall'informazione agli automobilisti (tramite pannelli luminosi, radio, internet ecc.) a interventi volti a ridurre il traffico, come ad esempio la chiusura o l'apertura di determinati segmenti stradali, il suggerimento di itinerari alternativi o l'adozione di limiti di velocità dinamici.

Allo stesso tempo, le piattaforme che forniscono servizi di navigazione agli automobilisti hanno forte interesse a integrare nei loro strumenti la capacità di prevedere la congestione. Ciò permetterebbe una migliore stima dei tempi d'itinerario, prendendo in considerazione le condizioni del traffico in tempo reale, e quindi un'ottimizzazione del servizio di navigazione offerto nei momenti critici della giornata.

Per tentare di prevedere un fenomeno complesso come la congestione stradale è necessario prima di tutto conoscerlo e quindi disporre di buoni dati di input. Qualunque sia la natura dei dati, questi devono essere affidabili, facilmente accessibili e disponibili in tempo reale. I

progressi nel campo dell'informatica e della telecomunicazione, insieme alla crescita esponenziale della potenza di calcolo, hanno reso disponibili delle grandi basi di dati. Questi dati, chiamati in inglese "*big data*", sono attualmente molto utilizzati in svariati ambiti a fini commerciali per proporre dei servizi personalizzati per ogni utente. [2]

La branca della scienza che si occupa dell'analisi e del trattamento dei big data si chiama "*data science*". Per trattare i big data è necessario utilizzare strumenti specifici in grado di gestire grandi volumi di dati e algoritmi che a partire da questi permettano di ottenere dei modelli di stima o previsione. La capacità di comprendere, interpretare trattare la mole di dati disponibili rappresenta una delle chiavi del successo del lavoro di un ingegnere dei trasporti.

L'obiettivo di questa tesi è proprio lo sviluppo di uno strumento che permetta di prevedere, a partire dai dati di mobilità e di traffico stradale attualmente disponibili, il volume di traffico e quindi la congestione in punti strategici della rete stradale, a un orizzonte temporale di 30 minuti. Questo strumento deve poter essere performante sia su una rete stradale urbana sia su una extraurbana. L'approccio scelto è quello del machine learning: si tratta di una tecnica che sfrutta l'intelligenza artificiale della macchina che viene sottoposta ad un processo di autoapprendimento. Il calcolatore si "allena" sulle serie storiche disponibili, analizzandole e identificando le relazioni statistiche tra le varie misure, arrivando a determinare un modello capace di prevedere ciò che accadrà in futuro a partire dai dati del presente. Il linguaggio di programmazione utilizzato è Python, linguaggio che si sta affermando come uno dei più apprezzati al mondo e che è già oggi il più utilizzato nel campo della data science. [3]

In questa tesi si cercherà di mettere in evidenza gli aspetti più inerenti al campo di studi di interesse, e quindi ci si concentrerà sugli elementi di ingegneria dei trasporti e sul contributo che l'ingegnere dei trasporti può apportare per rendere questo modello applicabile con precisione, affidabilità ed efficacia.

Lo studio è stato svolto presso gli uffici di Setec International a Parigi e si è ispirato ad una precedente ricerca realizzata nell'ambito del progetto Datacity. [4]

II. Contesto

In questo capitolo viene presentato il contesto in cui è nata e si è sviluppata la presente tesi nonché i progetti precedenti che l'hanno ispirata.

A. L'esperienza di Datacity

L'urbanizzazione è un fenomeno sempre più attuale nelle città. Lo sviluppo di centri abitati di grandi dimensioni comporta indubbiamente dei vantaggi dal punto di vista economico, ma porta con sé importanti problematiche dal punto di vista sociale, urbanistico ed ecologico. La pianificazione e la gestione di metropoli sempre più densamente popolate rappresentano delle sfide impegnative per le collettività. La diffusione in open source di un enorme volume di dati, i "*big data*", costituisce una risorsa fondamentale per individuare soluzioni capaci di rispondere a queste sfide. L'utilizzo di questa nuova risorsa per risolvere le problematiche urbane ha dato origine al concetto di "*smart city*", o città intelligente. La definizione di smart city non è univoca ed è attualmente oggetto di dibattito. [5] Una definizione semplificata, ma al tempo stesso chiara e immediata, è quella di Wikipedia France: "una città che integra le tecnologie dell'informazione e della comunicazione (TIC) e diversi dispositivi fisici connessi alla rete (l'Internet degli oggetti o IoT) per ottimizzare l'efficacia delle operazioni e dei servizi urbani e connettersi ai cittadini". [6]

Il progetto Datacity, promosso dal comune di Parigi, ha avuto come obiettivo proprio la ricerca di soluzioni tecnologiche e innovative ai problemi delle grandi metropoli. Uno degli ambiti di ricerca della terza edizione di questo progetto era la previsione, in tempo reale, della congestione stradale su uno specifico asse stradale[1]. I dati di input, acquistati e messi a disposizione dalla DiRIF[2], consistevano in velocità FCD (Floating Car Data), ossia velocità medie rilevate tramite GPS.

[1] A questa ricerca hanno partecipato Quantcube, start-up specializzata nel trattamento dei big data, la DiRIF e Setec International.

[2] Direction des Routes Ile-de-France: società concessionaria delle strade statali della regione parigina.

I risultati hanno evidenziato le potenzialità dell'utilizzo della data science nella previsione della congestione stradale. Al tempo stesso però ci si è resi conto che le prestazioni dello strumento non erano pienamente soddisfacenti, probabilmente a causa dell'approccio utilizzato durante Datacity. Infatti, a causa delle limitate risorse temporali disponibili, ci si è concentrati immediatamente sulla modellizzazione numerica, senza una preventiva fase di studio teorico del problema né un'analisi dei dati disponibili.

B. Lo sviluppo del progetto

I risultati incoraggianti ottenuti durante il progetto hanno comunque indotto a continuare la ricerca in questo campo, al fine di creare uno strumento che potesse essere operativo e affidabile. Un ulteriore obiettivo è stato provare ad allargare la previsione anche all'ambito urbano. Lo scopo è la realizzazione di uno strumento che, in un futuro non troppo lontano, possa divenire parte integrante delle competenze offerte da una società di ingegneria specializzata.

Lo sviluppo di un modello per la previsione del traffico necessita ovviamente di essere testato su casi concreti, che permettano di valutarne l'effettivo funzionamento. Per questo progetto, erano disponibili delle serie storiche di rilevazioni della rete extra-urbana della DiRIF, in quanto consulente del progetto Datacity. L'applicazione del modello all'ambito extra-urbano è stata quindi testata sulla rete stradale della regione parigina. L'applicazione del modello all'ambito urbano è stata invece testata sulla rete cittadina di Lione, terza città francese per numero di abitanti. In ambito urbano, le serie storiche dei dati di traffico erano disponibili in open data sul sito del comune.

Il progetto sviluppato nella mia tesi consiste nella rielaborazione dei dati di input utilizzati nell'esperienza di Datacity e nella costruzione ex-novo di un modello predittivo sulla base dei principi teorici dell'ingegneria dei trasporti e della data science.

III. Modelli predittivi e dati di input

L'obiettivo di questo capitolo è di introdurre i concetti teorici dell'ingegneria del traffico utilizzati per questo progetto, di descrivere i principali metodi di modellizzazione del traffico oggi utilizzati e di presentare i dati di input disponibili.

A. Cenni d'ingegneria del traffico stradale

Per analizzare e descrivere il traffico stradale, l'ingegneria dei trasporti propone due approcci comunemente adottati anche nella pratica professionale:

- **approccio microscopico:** ogni veicolo è modellato individualmente e il flusso di traffico è descritto determinando le dinamiche di interazione tra i singoli veicoli.
- **approccio macroscopico:** i singoli veicoli non vengono distinti e si parla di flusso di veicoli. Le variabili utilizzate per descrivere il traffico dal punto di vista macroscopico sono derivate dalla dinamica dei fluidi, in particolare portata o flusso, densità e velocità.

Per questo studio, il tipo di dati disponibili e il perimetro delle aree su cui il modello verrà testato suggeriscono un approccio di tipo macroscopico.

Questo approccio distingue due diverse condizioni di traffico:

- **flusso ininterrotto:** il flusso è influenzato esclusivamente dai veicoli al suo interno. Questo regime è tipico di assi stradali senza intersezioni, come le autostrade e le strade extraurbane principali.
- **flusso interrotto:** il flusso è influenzato da ostacoli esterni, come intersezioni a raso, restringimenti di carreggiata o semafori. Questo regime è frequente nelle reti stradali urbane.

1. Variabili macroscopiche

Come anticipato in precedenza, la natura e la dimensione dei dati disponibili, insieme al perimetro di interesse, hanno condotto alla scelta di un approccio di tipo macroscopico. Le variabili descrittive del traffico utilizzate durante questo progetto sono quindi:

- **portata:** è definita come il numero di veicoli che attraversano una sezione durante un intervallo di tempo t. Si esprime in veicoli/ora ed è indicata come Q. Nella maggior parte dei casi di osservazione continua del deflusso è misurata attraverso sensori, comunemente basati su spire induttive.

- **tasso d'occupazione:** è definito come la parte di tempo durante la quale il sensore è occupato da veicoli. Si esprime in percentuale ed è indicato come τ. Anch'esso nella maggior parte dei casi è misurato attraverso spire induttive.

- **densità:** è definita come il numero di veicoli presenti su una sezione stradale in un determinato istante. Si misura in veicoli/kilometro ed è indicata come K. La densità può essere ricavata dal tasso di occupazione tramite l'espressione:

$$K = \frac{0.001 \cdot \tau}{(l_b + \overline{l_v})}$$

  Dove $l_b[m]$ è la lunghezza della spira induttiva e $\overline{l_v}[m]$ è la lunghezza media dei veicoli. [7]

- **velocità media:** è definita come la velocità media del flusso di veicoli. Si misura in kilometri/ora ed è indicata con V. Può essere calcolata:

  - come media pesata delle singole misure di velocità, in cui il peso è il tempo impiegato dai singoli veicoli nel dominio di misura. Si parla in questo caso di velocità media temporale. Le singole velocità possono essere rilevate con appositi sensori (ad esempio tramite radar o GPS).

  - a partire dalla portata e dalla densità tramite l'espressione:

$$V = \frac{Q\ (l_b + \overline{l_v})}{0.001 \cdot \tau}$$

2. Diagrammi fondamentali

Le relazioni tra le variabili appena introdotte sono state ampiamente studiate in passato e in letteratura son riportati molti modelli. Le prime curve portata-velocità sono state determinate da Greenshields e successivamente affinate da Greenberg. [8] [9] [10] Queste curve prendono il nome di diagrammi fondamentali.

La Figura 1 illustra il legame tra portata e densità. La relazione è una sorta di parabola asimmetrica che presenta un massimo in corrispondenza di un preciso valore di densità, detto densità critica. Questa curva mostra che a basso valore di portata può corrispondere un basso valore di densità (regime di traffico non congestionato) o un grande valore di densità (regime di traffico congestionato). È quindi evidente che la sola misura della portata non è sufficiente per descrivere le condizioni del traffico.

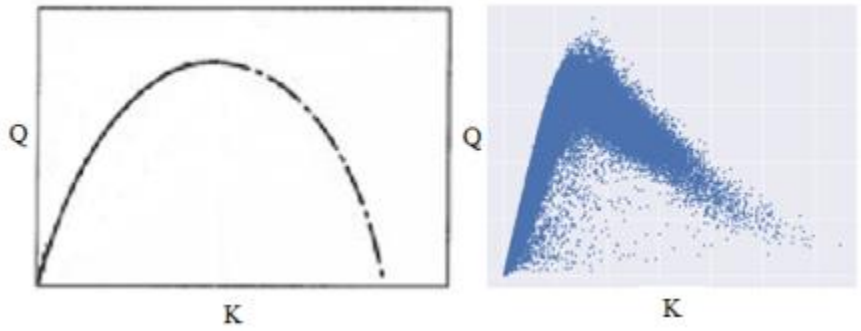


Figura 1: Diagramma fondamentale portata-densità teorico (secondo Greenshields) e sperimentale (ricavato dalle misure della DiRIF)

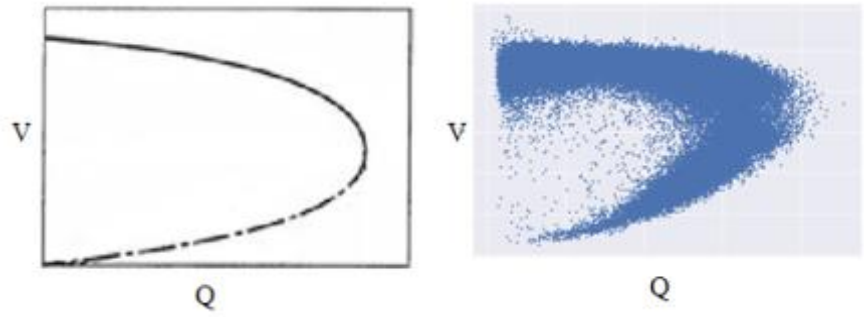


Figura 2: Diagramma fondamentale portata-velocità teorico (secondo Greenshields) e sperimentale (ricavato dalle misure della DiRIF)

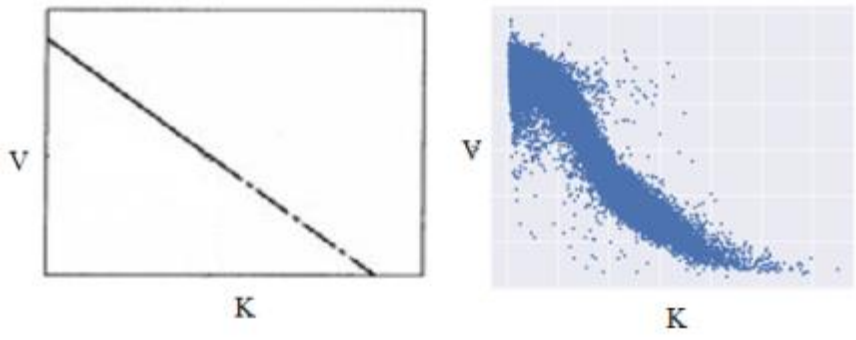


Figura 3: Diagramma fondamentale velocità-densità teorico (secondo Greenshields) e sperimentale (ricavato dalle misure della DiRIF)

B. Modelli predittivi

Come anticipato, la tecnica scelta per la previsione della congestione è quella del machine learning. Questa tecnica è anche detta *data-driven*, in quanto permette di prevedere le condizioni del sistema nel futuro a partire dall'analisi delle serie temporali. Gli algoritmi di machine learning analizzano i dati storici delle serie temporali per dedurne delle relazioni statistiche che utilizzano per prevedere, a partire dalle condizioni di traffico attuali, quelle future. In letteratura è consolidato un altro approccio di previsione, che potremmo definire più classico, detto model-driven.

1. Previsione model-driven

I modelli predittivi detti *model-driven*, letteralmente "previsione guidata dal modello", si basano su modelli fisico-matematici di traffico. Questi modelli sono derivati a loro volta dalla teoria dell'ingegneria del traffico. [11] Secondo la scala considerata si parla di modelli macroscopici o microscopici. [12]

L'approccio macroscopico è attualmente utilizzato per studiare l'effetto della congestione su una rete stradale di grandi dimensioni attraverso diversi modelli di simulazione. Questo processo necessita di un livello di dettaglio molto elevato in termine di dati di input. In più l'utilizzo di questi modelli richiede la conoscenza dello stato iniziale nel sistema e del profilo della domanda nelle sezioni di ingresso [13]; le caratteristiche principali che vengono richieste ai dati di input sono la precisione e l'accuratezza.

I modelli microscopici si posizionano alla scala del veicolo isolato e ricostruiscono le condizioni del traffico valutando il comportamento del singolo automobilista. Attraverso tale approccio, simulando la traiettoria e la velocità dei veicoli, è possibile confrontare diversi scenari; per tale ragione i modelli microscopici sono spesso utilizzati per valutare le strategie di regolazione del traffico. Il moto dei veicoli è simulato con modelli following-car (legge dell'inseguitore) [14] e modelli a cambio di corsia [2].

2. Previsione data-driven

La previsione *data-driven*, letteralmente "previsione guidata dai dati", è una tecnica che ha conosciuto una popolarità crescente negli ultimi 20 anni. La disponibilità dei big data, per definizione un enorme volume di dati, in tempo reale e il contemporaneo sviluppo esponenziale dell'intelligenza artificiale hanno permesso la creazione di algoritmi predittivi che si basano esclusivamente sulle relazioni statistiche tra le serie storiche, senza passare attraverso dei modelli matematici che tentano di spiegare il comportamento del sistema.

Il valore di una variabile ad un istante di tempo $t + n$ è:

$$y_{t+n} = f(X_{t,t-1\dots t-m})$$

Prendendo quindi le variabili X all'istante attuale t e agli istanti passati $t - 1, \dots, t - m$ come variabili esplicative, è possibile esplicitare la variabile y all'istante $t + n$.

L'espressione *machine learning*, letteralmente "apprendimento della macchina", raggruppa due metodi: l'apprendimento supervisionato (*supervised learning*) e l'apprendimento non supervisionato (*unsupervised learning* o *clustering*). [15] Entrambi i metodi sono stati utilizzati, in momenti diversi, in questo studio. Si rende a questo punto necessaria una breve spiegazione dei loro principi di funzionamento.

- **L'apprendimento non supervisionato** o clustering consiste nel raggruppare gli individui di una popolazione in classi analizzando le relazioni tra le loro variabili descrittive e minimizzando la distanza interclasse. Una volta che i dati sono preparati per il trattamento e forniti all'algoritmo, questo classifica gli individui in modo autonomo (e quindi non supervisionato) senza alcuna necessità né possibilità di intervento dall'esterno. Il clustering è utilizzato per classificare degli elementi le cui categorie di appartenenza non sono note a priori.

- **L'apprendimento supervisionato**, al contrario, analizza i dati storici, la cui classificazione è nota, per dedurre le caratteristiche che individuano ciascuna classe. Una volta che l'algoritmo ha potuto "allenarsi" sui dati storici (definendo la funzione $f(x)$ modellata sulle relazioni statistiche delle serie storiche), è pronto per classificare i nuovi individui.

C. Dati di input

Gli strumenti tecnologici che permettono di misurare le variabili introdotte in precedenza possono avere forme e principi di funzionamento anche molto diversi. Per questo progetto i dati di input provengono sia da spire induttive sia da FCD. Nel seguito sono descritti i principi di funzionamento di entrambi gli strumenti di misura. Inoltre, i due tipi di dati saranno confrontati tra loro per valutarne l'affidabilità.

1. Spire induttive

La spira induttiva è il tipo di sensore più conosciuto e, soprattutto in passato, il più diffuso. Questi rilevatori permettono una misura precisa e affidabile della portata e del tasso di occupazione.

Inventata negli anni Sessanta [16], la spira induttiva è un elemento che si inserisce nella pavimentazione stradale. Una componente elettronica trasmette una corrente elettrica ad una frequenza compresa tra 10 e 200 kHz. Il passaggio del veicolo induce una corrente di Foucault all'interno della spira e tale corrente produce un abbassamento della frequenza. La spira è collegata tramite un cavo ad un rilevatore di frequenza: al passaggio del veicolo questo capta l'abbassamento di frequenza e invia un segnale all'apparecchio di controllo indicando la presenza di una vettura sulla spira. [17] [18]

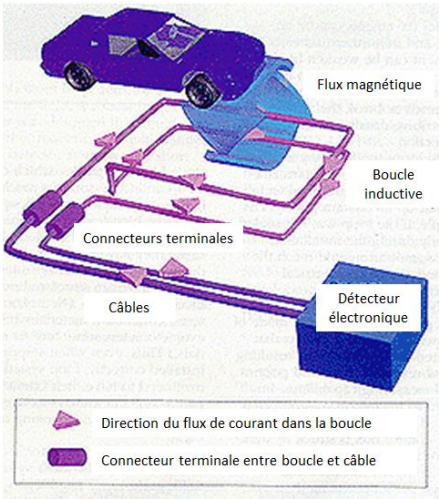


Figura 4: Schema del funzionamento di una spira induttiva

Fonte: Federal Highway Administration

La spira induttiva è un sensore fisicamente fissato all'infrastruttura, che permette una misura puntuale e fissa. Si tratta di un sistema intrusivo che necessita di interventi d'ingegneria per essere installato e gestito. In effetti, il principale difetto di questo tipo di rilevatori è il costo d'installazione e soprattutto di manutenzione molto elevato. In più, qualsiasi rinnovamento o intervento sulla pavimentazione stradale risulta complicato dalla presenza della spira. Ciononostante, le misure delle spire sono affidabili, precise e normalmente disponibili in continuo. In più questo tipo di sensore permette la misura (diretta o indiretta) di tutte le principali variabili macroscopiche, come portata, densità e velocità media del flusso.

Per entrambi gli ambiti di studio affrontati nella tesi, la rete urbana di Lione e quella extra-urbana della DiRIF, erano disponibili misure rilevate tramite spire induttive con valori di portata e di tasso di occupazione su intervalli di 6 minuti. Si trattava di dati grezzi, direttamente registrati dai sensori senza alcun post-trattamento né verifiche da parte dei fornitori. Soprattutto nel caso della città di Lione, la verifica preliminare condotta ha evidenziato il funzionamento parzialmente o totalmente compromesso di alcuni sensori (mancanza di più della metà dei dati, serie con alternanza di 0 e di 1, valori di portata negativi). Si è dunque scelto di non prendere in considerazione i dati provenienti dalle spire con funzionamento significativamente compromesso. Tale scelta non ha avuto ripercussioni significative sul progetto in quanto, una volta effettuata la scrematura su un totale di circa 570 sensori, il numero di quelli affidabili è rimasto comunque superiore a 500.

Per Lione i dati erano disponibili in open source [19] mentre per la rete DiRIF sono stati forniti a Setec International in quanto consulente del progetto Datacity.

2. Floating Car Data

L'acronimo FCD (*Floating Car Data*) identifica dei dati di navigazione GNSS estratti dagli apparecchi montati a bordo dei veicoli. Si tratta della velocità media[3] della flotta di veicoli

[3] La velocità media è calcolata dalle compagnie di telecomunicazione *prima* che le FCD vengano commercializzate. Per tale ragione non è stato possibile, nel caso di questo progetto, verificare se i valori a disposizione rappresentino delle velocità medie spaziali o delle velocità medie temporali.

presente in un determinato momento su un segmento stradale. In realtà non si tratta della velocità media di tutti i veicoli, ma solo di quelli equipaggiati con GPS attivo.

Il principio di funzionamento delle FCD è illustrato in Figura 5. I recettori GNSS dei veicoli captano il segnale emesso dai satelliti (1) e lo trasformano in un'informazione di posizione, direzione e velocità (2). Queste informazioni sono trasmesse ad un ricevitore, associate a un *timestamp* e all'identificativo del veicolo (3). I dati sono elaborati da server di proprietà delle compagnie di telecomunicazione, che ne calcolano ad esempio la velocità media (4). Dopo il post-trattamento i dati vengono commercializzati (5).

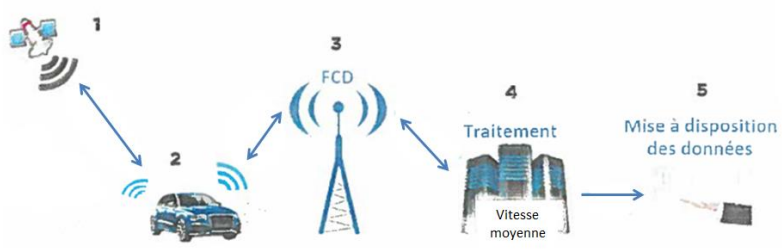


Figura 5: Schema del funzionamento dei dati FCD

Fonte: [20]

I dati FCD provengono direttamente da rilevatori integrati nelle apparecchiature montate sui veicoli degli utenti e di conseguenza il loro unico costo è quello associato al post-trattamento. Non è necessaria alcuna installazione di materiale nella pavimentazione. Indicativamente, la manutenzione e la modernizzazione di un sensore tradizionale (spira induttiva) costa 15 000 € mentre il costo dell'acquisto di dati FCD, per un anno su un tratto di 1 km, è di circa 100 €. [20] L'ulteriore vantaggio di questa tecnologia, insieme al basso costo, è la possibilità di avere misure relative a tutti i punti della rete stradale (e non solo nei punti dove è stato installato un sensore).

I limiti dei dati FCD sono legati alla natura stessa della tecnologia: trattandosi di una velocità calcolata come media a partire dai veicoli presenti su un segmento stradale, se la portata è debole questa misura può non essere affidabile o addirittura non disponibile. Inoltre, le misure FCD non sono disponibili in galleria a causa dell'assenza del segnale GNSS e la precisione è limitata alla carreggiata anziché alla corsia (a causa dell'attuale precisione del GPS). Infine, non si ha alcuna informazione né sulla portata né sulla densità del flusso veicolare.

Nei casi studiati, i dati FCD erano disponibili per la rete DiRIF ma non per la città di Lione. Le velocità medie della rete DiRIF erano aggregate per intervalli di 6 minuti, quindi perfettamente confrontabili con i dati delle spire induttive. Si trattava di dati che avevano subito un post-trattamento da parte degli operatori di telefonia che li avevano raccolti. Il confronto tra i dati delle spire e i dati FCD, esposto nel paragrafo seguente, ha permesso di valutare l'affidabilità di queste misure.

3. Confronto tra dati FCD e dati delle spire induttive

Le velocità FCD rappresentano, grazie alle caratteristiche appena descritte, un tipo di dato destinato ad essere sempre più utilizzato in futuro. Per questa ragione è necessario avere una buona conoscenza delle caratteristiche e dell'affidabilità di questo nuovo strumento. A tal fine, prima di utilizzare questi dati nel modello è stato condotto un confronto tra le velocità calcolate a partire dal diagramma fondamentale e quelle misurate tramite le FCD.

Come già accennato, nel caso della rete extra-urbana parigina erano disponibili sia i dati rilevati dalle spire induttive sia i dati FCD. Entrambi i dati erano disponibili per intervalli di 6 minuti e quindi comparabili. Una giornata risultava dunque divisa in 240 intervalli, per un totale di 175 440 intervalli per ogni punto di rilevamento sui due anni considerati, 2016 e 2017.

La convenzione utilizzata dalla DiRIF per individuare i segmenti nei due sensi di marcia è la seguente:

- E/XX+XXXX per la direzione "*extérieure*" (senso antiorario intorno a Parigi)
- I/XX+XXXX per la direzione "*intérieure*" (senso orario intorno a Parigi)

Il segmento stradale scelto per il confronto è quello compreso tra la progressiva 40+0000 e la progressiva 45+0000 della strada nazionale N104 della rete DiRIF: tale scelta è dettata dall'importanza di quest'asse (chiamato *la Francilienne*) tra quelli per cui i dati erano disponibili. L'arteria esaminata è una strada extra-urbana nazionale a carreggiate separate, con due corsie per senso di marcia e limite di velocità di 90 km/h su tutto il tratto studiato. In direzione *intérieure* sono presenti sei intersezioni (tre bretelle d'ingresso e tre d'uscita), in direzione *extérieure* sono presenti nove intersezioni (quattro bretelle d'ingresso e cinque

d'uscita). I dati FCD erano disponibili per tratti stradali di 200 metri, quindi 25 segmenti per i 5 km considerati. Su alcuni di questi tratti erano presenti anche delle spire induttive, permettendo così di disporre di dati accoppiati. Dati FCD supplementari erano inoltre disponibili laddove sono presenti delle intersezioni. Per il confronto sono stati presi in considerazione solo i 14 tratti stradali su cui era presente anche una spira.

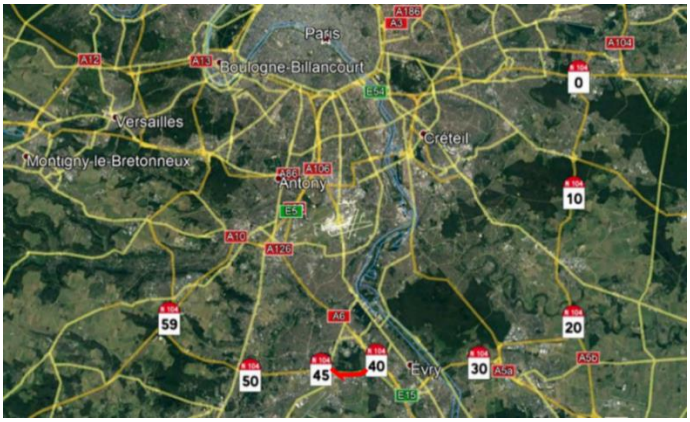


Figura 6: Localizzazione del segmento stradale della N104 tra la PR 40+0000 e la PR 45+0000
Fonte: Google Earth

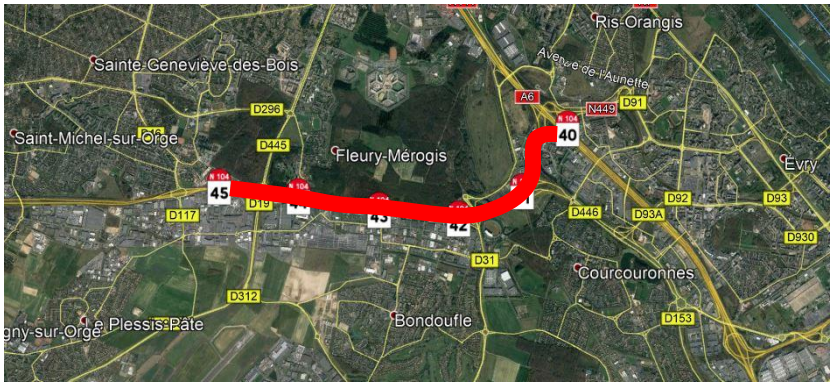


Figura 7: Dettaglio del segmento stradale della N104 tra la PR 40+0000 e la PR 45+0000
Fonte: Google Earth

a. Disponibilità dei dati FCD

Il primo test effettuato aveva l'obiettivo di verificare la disponibilità dei dati FCD in funzione dell'ora e del giorno della settimana. Dal momento che, per questo tipo di dati forniti da terzi, si disponeva unicamente della velocità media, senza informazioni supplementari sul numero di veicoli a partire dai quali tale valore è stato calcolato, con il termine disponibilità si intende semplicemente l'esistenza della misura sull'intervallo considerato. Come accennato in

precedenza, la disponibilità delle FCD non è sempre garantita ed è legata alle condizioni di traffico. La disponibilità è calcolata come il rapporto (in percentuale) tra il numero di intervalli di 6 minuti per cui le rilevazioni sono disponibili e il numero totale di intervalli durante i due anni.

Sul segmento stradale considerato, i risultati in Tabella 1 mostrano una buona disponibilità dei dati FCD: per quasi tutta la giornata infatti si ha una disponibilità media superiore all'80%. Solamente durante le ore notturne dei giorni lavorativi e la domenica mattina la media della disponibilità scende al disotto del 75%. Come ci si aspettava tale diminuzione si verifica nei momenti in cui il traffico è meno intenso. Tenendo conto che l'obiettivo del progetto è la previsione della congestione, questi risultati confermano che i dati FCD sono generalmente disponibili: infatti, nel momento in cui il numero di veicoli aumenta, caratteristica propria della congestione, le FCD sono sempre disponibili.

	Lunedì	Martedì	Mercoledì	Giovedì	Venerdì	Sabato	Domenica
00:00	75%	73%	74%	75%	80%	85%	83%
01:00	66%	67%	67%	68%	74%	84%	85%
02:00	59%	62%	62%	63%	70%	80%	79%
03:00	66%	69%	69%	69%	75%	79%	73%
04:00	77%	77%	76%	77%	82%	81%	69%
05:00	76%	77%	75%	76%	80%	80%	65%
06:00	79%	78%	78%	78%	82%	82%	65%
07:00	81%	81%	78%	79%	83%	81%	70%
08:00	83%	82%	79%	81%	84%	86%	76%
09:00	83%	81%	79%	82%	85%	87%	82%
10:00	83%	81%	80%	82%	85%	87%	86%
11:00	83%	80%	79%	81%	85%	87%	86%
12:00	84%	82%	80%	83%	86%	87%	86%
13:00	83%	82%	81%	84%	86%	87%	86%
14:00	82%	81%	81%	83%	86%	88%	87%
15:00	82%	81%	81%	81%	86%	88%	86%
16:00	83%	83%	81%	83%	87%	88%	86%
17:00	84%	83%	82%	84%	87%	88%	86%
18:00	84%	83%	83%	85%	88%	88%	86%
19:00	84%	84%	83%	85%	88%	86%	85%
20:00	83%	82%	82%	83%	86%	87%	85%
21:00	81%	81%	81%	82%	86%	84%	82%
22:00	81%	80%	80%	82%	86%	85%	83%
23:00	80%	80%	80%	82%	86%	86%	83%

Tabella 1: Disponibilità delle FCD per ora e per giorno della settimana[4]

La Tabella 2 rappresenta invece il minimo della disponibilità sui 14 tratti stradali presi in esame. Si può osservare che per le ore centrali dei giorni feriali la disponibilità dei dati è sempre superiore al 60%, e al 65% nella maggior parte dei casi.

[4] La disponibilità è calcolata come media sui 14 segmenti del tratto stradale considerato. Si tratta dei 14 segmenti per cui le FCD erano accoppiate a una rilevazione tramite spira induttiva.

	Lunedì	Martedì	Mercoledì	Giovedì	Venerdì	Sabato	Domenica
00:00	52%	46%	50%	56%	62%	71%	64%
01:00	28%	28%	30%	33%	41%	64%	63%
02:00	13%	19%	21%	22%	29%	56%	46%
03:00	19%	31%	32%	33%	41%	52%	38%
04:00	51%	56%	57%	53%	61%	47%	29%
05:00	50%	51%	49%	47%	53%	60%	22%
06:00	57%	57%	57%	56%	57%	53%	20%
07:00	61%	64%	62%	60%	64%	47%	11%
08:00	71%	72%	67%	68%	71%	69%	15%
09:00	71%	70%	67%	68%	72%	73%	32%
10:00	70%	70%	67%	68%	67%	73%	66%
11:00	70%	70%	67%	69%	71%	72%	71%
12:00	71%	70%	67%	69%	72%	72%	71%
13:00	70%	70%	68%	69%	72%	72%	72%
14:00	71%	71%	68%	69%	72%	72%	72%
15:00	70%	70%	68%	68%	72%	73%	71%
16:00	70%	70%	68%	69%	73%	73%	71%
17:00	71%	70%	68%	69%	73%	73%	72%
18:00	72%	70%	69%	70%	73%	72%	71%
19:00	72%	70%	69%	70%	73%	72%	72%
20:00	72%	71%	69%	70%	72%	72%	63%
21:00	66%	66%	65%	69%	73%	51%	41%
22:00	69%	67%	67%	69%	72%	72%	59%
23:00	66%	66%	67%	68%	72%	71%	61%

Tabella 2: Minimo della disponibilità delle FCD per ora e per giorno della settimana

Prendendo come disponibilità limite il 70%, la Tabella 3 mostra che ad eccezione delle ore notturne e della domenica mattina, l'informazione è sempre disponibile per almeno 10 segmenti su 14.

	Lunedì	Martedì	Mercoledì	Giovedì	Venerdì	Sabato	Domenica
00:00	4	4	4	4	2	0	1
01:00	7	7	7	7	4	1	2
02:00	9	7	8	7	6	3	3
03:00	7	6	6	5	3	2	4
04:00	3	3	4	4	2	2	4
05:00	3	3	4	4	1	2	6
06:00	2	2	3	3	1	2	5
07:00	1	1	2	2	1	2	5
08:00	0	0	2	1	0	2	3
09:00	0	0	3	1	0	0	2
10:00	0	0	2	2	1	0	1
11:00	1	0	2	1	0	0	0
12:00	0	0	2	1	0	0	0
13:00	1	0	1	1	0	0	0
14:00	0	0	1	1	0	0	0
15:00	0	0	2	1	0	0	0
16:00	0	0	1	1	0	0	0
17:00	0	0	1	1	0	0	0
18:00	0	0	1	1	0	0	0
19:00	0	0	1	1	0	0	0
20:00	0	0	1	1	0	0	1
21:00	1	1	2	2	0	2	2
22:00	1	2	3	1	0	0	1
23:00	1	2	3	2	0	0	2

Tabella 3: Numero dei tratti stradali per i quali la disponibilità delle FCD è inferiore al 70%

b. Affidabilità dei dati FCD

Per valutare la qualità dell'informazione fornita dalle velocità FCD, è stato realizzato un confronto con le velocità medie ricavate dalle misure delle spire induttive. La dimensione delle spire era nota grazie alle informazioni della DiRIF e, facendo un'ipotesi sulla lunghezza media dei veicoli, è stato possibile calcolare la velocità media del flusso.

Per questo confronto, non sono stati presi in considerazione i punti di accesso alla rete, che corrispondono ai sensori situati sulle bretelle di intersezione: in effetti la portata su queste bretelle è generalmente troppo debole per fornire dei dati FCD affidabili. Inoltre, in una prospettiva operativa, le previsioni del volume di traffico su queste bretelle hanno poco interesse pratico. La DiRIF ha infatti confermato come sui tratti di accesso alla rete l'informazione di portata sia molto più importante che altrove e che, in corrispondenza di tali punti, l'informazione proveniente dalle spire non potrà essere sostituita dalle FCD.

I risultati evidenziano due tipi di curve differenti:

- per alcuni segmenti, come in Figura 8, si osserva una corrispondenza quasi perfetta tra la velocità media misurata tramite i valori FCD (in ordinata) e quella calcolata a partire dalle misure delle spire (in ascissa). La maggior parte dei punti si trova sulla diagonale, il che significa che le due misure di velocità corrispondono. Per velocità più elevate (quindi per un regime non congestionato) le FCD sovrastimano leggermente la velocità, come peraltro rilevato anche dalla DiRIF [20]. Inoltre, qualche errore di sottostima può essere osservato nella parte bassa del diagramma. Ciononostante, la correlazione tra le misure resta ottima.

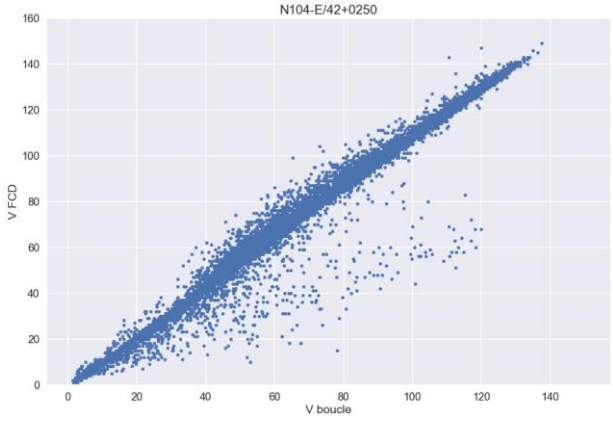


Figura 8: Esempio di sensore con un'ottima corrispondenza tra velocità FCD e velocità della spira

- per alcuni segmenti, come in Figura 9, si possono osservare dei punti con un comportamento significativamente differente: tali punti formano una nuvola nella parte superiore del diagramma. Mettendo in evidenza queste misure sul diagramma portata-densità (Figura 10), si osserva che queste si concentrano in corrispondenza di densità basse e per un largo spettro di valori di portata. Il diagramma presenta in ascissa il tasso di occupazione anziché la densità: dal momento che le due variabili sono direttamente proporzionali, si è preferita una rappresentazione in termini di tasso di occupazione perché tale grandezza non dipende dall'ipotesi sulla lunghezza media dei veicoli. Osservando la heatmap in Tabella 4, è evidente come questi punti con un comportamento differente si concentrino durante le ore notturne dei giorni feriali. I risultati sono espressi in numero totale di misure anomale per ogni fascia oraria. Al di là di queste evidenti anomalie, la correlazione resta buona come nel caso precedente.

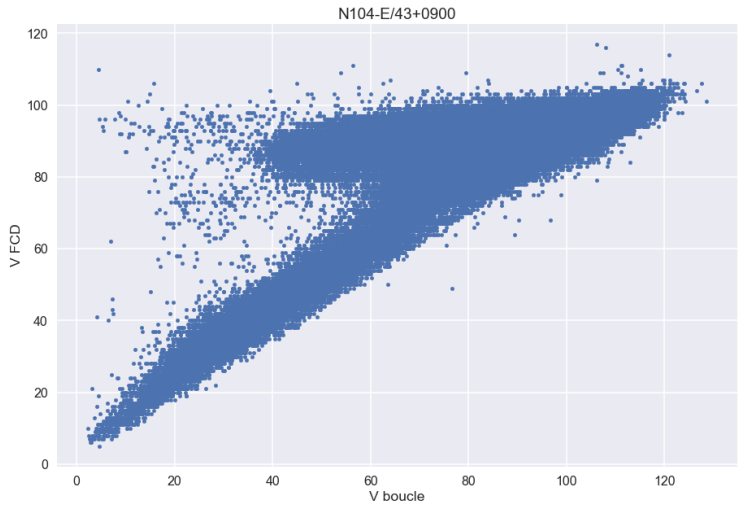


Figura 9: Esempio di sensore con delle anomalie di rilevazione

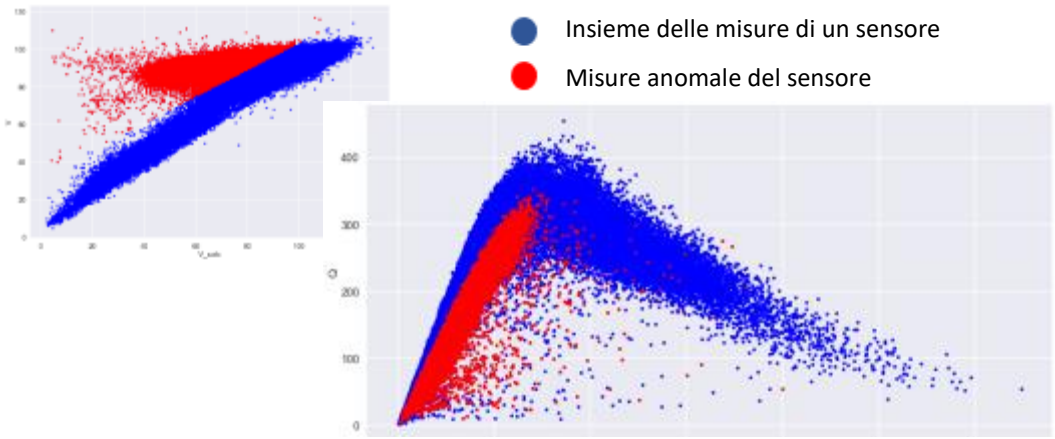


Figura 10: Localizzazione delle anomalie sul diagramma fondamentale

	Lunedì	Martedì	Mercoledì	Giovedì	Venerdì	Sabato	Domenica
00:00	216	550	558	551	520	126	0
01:00	447	633	663	642	648	439	0
02:00	558	652	669	652	679	530	0
03:00	635	661	686	669	693	630	1
04:00	650	669	677	658	688	669	0
05:00	476	496	470	458	460	552	4
06:00	455	511	509	490	490	425	11
07:00	36	110	109	105	86	167	2
08:00	0	2	3	0	0	16	2
09:00	21	37	54	36	22	10	4
10:00	109	226	206	194	67	8	5
11:00	81	168	129	120	23	9	6
12:00	34	47	29	37	6	0	4
13:00	17	21	26	16	3	2	1
14:00	6	12	7	3	4	2	1
15:00	7	20	17	6	1	1	0
16:00	3	0	0	0	4	2	0
17:00	0	0	0	4	1	0	2
18:00	3	0	0	1	0	0	0
19:00	7	3	1	1	0	0	3
20:00	20	5	11	4	2	0	1
21:00	185	138	137	93	3	0	3
22:00	226	264	233	161	28	0	5
23:00	379	369	333	302	34	0	29

Tabella 4: Heatmap della distribuzione delle anomalie

I due punti dell'asse stradale utilizzati per il confronto tra misure delle spire e misure FCD, rappresentati in Figura 8 e Figura 9, sono stati scelti a titolo d'esempio, ma il comportamento di tutti i punti dell'asse studiato può essere ricondotto a uno dei due casi appena illustrati. Rappresentando i punti sulla carta (Figura 11), non è possibile osservare nessuna relazione spaziale tra quelli con comportamento più "pulito" (in bianco) e quelli che presentano una nuvola di punti non correlati (in rosso).

Per una migliore comprensione di questo particolare comportamento delle FCD per alcuni punti sarebbe necessario condurre un'analisi più approfondita, che esula dagli obiettivi di questo studio. Tuttavia, è possibile ipotizzare alcuni fattori che potrebbero spiegare il fenomeno osservato. L'ipotesi più plausibile è che la velocità media FCD sia calcolata come media temporale, mentre la velocità media a partire dalle spire sia calcolata come media spaziale. Se così fosse, la correlazione tra le due misure dipenderebbe dal numero di veicoli sui quali la media è calcolata: per un flusso importante i due valori dovrebbero essere simili, per un flusso più debole la correlazione potrebbe diminuire. La Tabella 4 mostra come la correlazione sia effettivamente peggiore durante le ore notturne, quando la portata è più bassa: ciò sembrerebbe quindi confermare questa ipotesi. Tuttavia, l'assenza d'informazioni più precise sul metodo di calcolo della velocità media FCD non permette di confermare quanto appena ipotizzato.

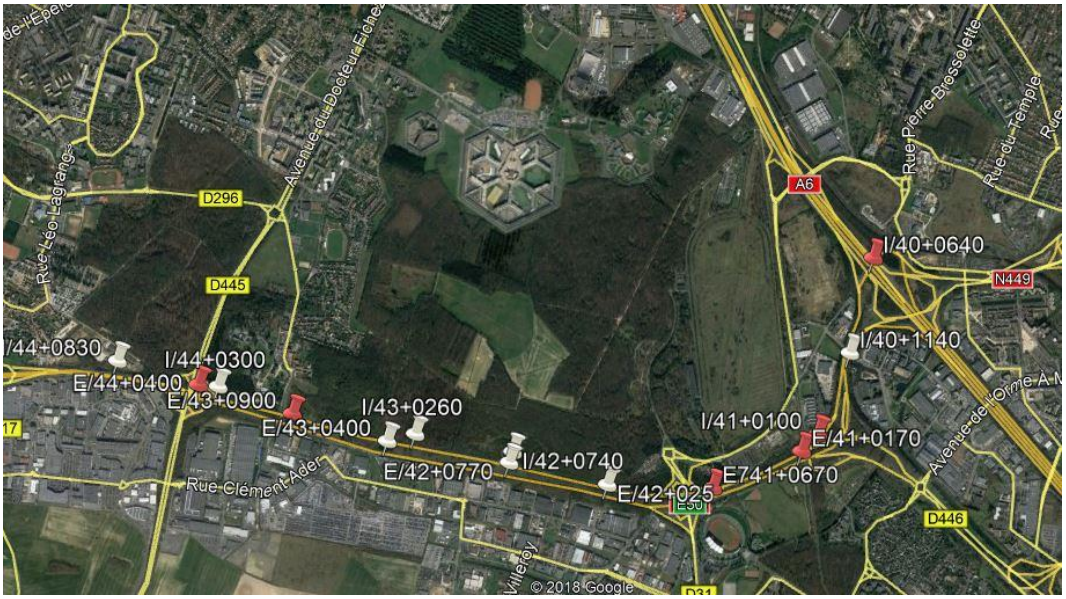


Figura 11: Localizzazione dei sensori con (in rosso) e senza (in bianco) anomalie

Fonte: Google Earth

Il confronto condotto in questo paragrafo ha permesso di evidenziare una buona disponibilità dei dati FCD ma allo stesso tempo ha mostrato che tali dati presentano dei comportamenti di difficile interpretazione per alcuni sensori e per alcune particolari situazioni di traffico.

Per tale ragione, e anche perché i dati FCD erano disponibili esclusivamente per la rete extra-urbana della DiRIF, si è scelto di alimentare il modello esclusivamente attraverso le rilevazioni di portata e densità delle spire induttive. Tali misure sono infatti ritenute più affidabili; esse erano inoltre disponibili sia per la rete DiRIF che per la città di Lione.

IV. Metodologia

L'obiettivo di questo capitolo è presentare la metodologia utilizzata nel corso del progetto. In linea generale si possono identificare quattro tappe fondamentali:

- Trattamento dei dati mancanti
- Manipolazione dei dati per trasformarli in variabili esplicative per il modello
- Previsione della portata e del tasso di occupazione
- Definizione della congestione

A. Trattamento dei dati mancanti

Dopo un'esplorazione preliminare dei dati disponibili, è stato possibile osservare che alcuni sensori fornivano delle informazioni parziali, con dei "buchi" importanti nelle rilevazioni (Figura 12). Questi buchi possono dipendere dal malfunzionamento momentaneo delle spire, da interventi di manutenzione della pavimentazione stradale o dall'indisponibilità temporanea della copertura GPS. La presenza di dati incompleti è stata riscontrata sia nel caso dei sensori lionesi sia nel caso delle spire e delle velocità FCD della DiRIF. Il trattamento dei dati mancanti nelle serie temporali è oggetto di ricerca e in letteratura sono proposte diverse soluzioni. [21] [22]

In questo progetto i buchi all'inizio e alla fine delle serie sono semplicemente stati scartati accorciando il periodo di osservazione. D'altro canto, l'aver accorciato tali periodi non influenza i risultati finali data l'ampia durata delle rilevazioni (due anni per la rete DiRIF e tre anni per la rete lionese). Cionondimeno la scelta di un approccio meno "semplificato" costituisce una delle possibili piste di miglioramento del metodo. Al contrario, i buchi in mezzo ad una serie sono stati trattati, in quanto gli algoritmi di machine learning non possono applicarsi a delle serie con valori mancanti. Tali valori, individuati in Python come NaN (*Not a Number*), sono stati stimati attraverso un'interpolazione lineare.

Si tratta di una soluzione rapida e di facile applicazione, anche se comporta l'introduzione di approssimazioni dei risultati. In ogni caso, la sperimentazione è stata condotta su dei sensori e

dei segmenti stradali specificamente scelti per l'elevata qualità dei dati e ciò ha permesso di limitare il numero di casi in cui è stato necessario intervenire.

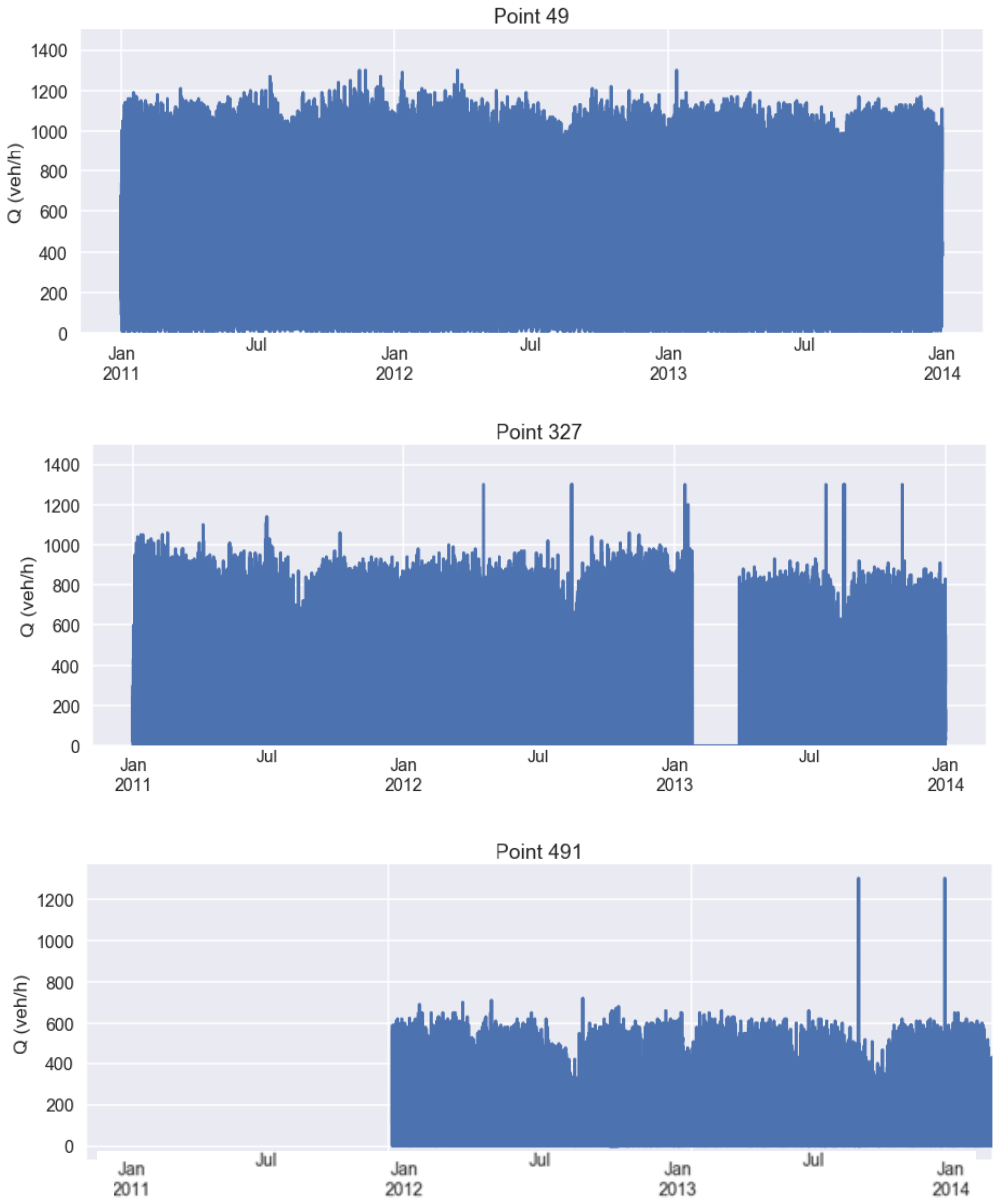


Figura 12: Esempio di sensori con serie temporali con differenti gradi di dati mancanti – Misure di portata delle spire induttive di Lione

B. Creazione delle variabili esplicative del modello

Il traffico è un fenomeno complesso e sono necessarie variabili specifiche per modellizzarlo e tentare di prevederlo. In primo luogo ha un carattere ciclico e deve essere presa in considerazione una sua componente temporale. Inoltre, il traffico in un punto della rete stradale dipende anche da quello presente sui segmenti prossimi al punto in esame negli istanti precedenti: di conseguenza deve essere considerata anche una componente spaziale. Infine, sono stati considerati altri elementi che possono influenzarlo, come ad esempio le condizioni meteo.

1. Aspetti temporali

Come appena accennato, il traffico ha un comportamento ciclico nel tempo. È possibile identificare dei ritmi giornalieri, settimanali e stagionali. L'introduzione di tale aspetto nella modellizzazione può semplificare di molto la previsione: se una parte del comportamento della serie può essere spiegata attraverso la ripetizione di un ritmo settimanale noto e costante, allora il modello dovrà prevedere solo la varianza rispetto a tale tendenza.

a. Stagionalità

La libreria *statsmodel* di Python mette a disposizione un modulo chiamato *seasonal_decompose* che permette, a partire da una serie temporale e con una frequenza scelta dall'utente, di disaggregare la stessa in tre "sotto-serie": un *trend* T, una *stagionalità* S e un *residuo* R. [23] Questa disaggregazione, detta additiva, si esprime come:

$$y(t) = T(t) + S(t) + R(t)$$

Per considerare le abitudini settimanali degli automobilisti, è stata scelta una frequenza di sette giorni. Ogni settimana è quindi formata da 1680 misure (misure su intervalli di 6 minuti).

- **il trend** considera delle tendenze (crescenti o decrescenti) dei valori della serie. Nel modulo *seasonal_decompose* il trend è in realtà un trend-ciclo, ovvero contiene sia delle tendenze vere e proprie sia delle oscillazioni cicliche dei valori con una frequenza diversa da quella scelta. Dal punto di vista operativo il trend è calcolato attraverso una

media mobile centrata in cui l'ampiezza della finestra mobile coincide con la frequenza della stagionalità.

- **la "stagionalità"** (traduzione dell'inglese *seasonnality*) esprime le oscillazioni che si verificano con una frequenza prestabilita. Secondo la definizione di frequenza scelta, nella stagionalità sono presi in considerazione i ritmi settimanali. In particolare, la stagionalità fornita dal modulo *seasonal_decompose* è un'oscillazione rispetto al trend ed è calcolata come media di tutte le misure dello stesso periodo (stesso giorno della settimana e stessa fascia oraria). Proprio per la natura della sua definizione, la stagionalità è un valore relativo rispetto al trend. Per ovviare a questo problema, si è scelto di calcolare un valor medio del trend su tutto il periodo di osservazione e sommarlo alla stagionalità precedentemente calcolata che risulterà in questo modo de-relativizzata. La perdita di informazioni dovuta alla semplificazione del trend, che viene sostituito dal suo valor medio, non introduce un errore sistematico: infatti tali informazioni entreranno a far parte dei valori che si chiederà al modello di prevedere. Nel paragrafo IV.B.1.b, verranno introdotte delle variabili supplementari per tenere conto degli effetti del calendario.

- **il residuo** rappresenta tutte le variazioni che non sono esplicate dai due elementi precedenti. É esattamente qui che sono nascoste le relazioni, e dunque le informazioni, che l'algoritmo deve esplorare e comprendere per poter prevedere la congestione. A causa della modifica apportata alla stagionalità e al trend, ciò che verrà definito residuo nel seguito sarà la differenza tra il valore osservato e la stagionalità de-relativizzata.

Come illustrato in Figura 13, che riporta l'esempio del sensore 226 di Lione durante una settimana del gennaio 2013, la serie temporale del flusso misurato viene disaggregata e solo la parte di residuo necessita una previsione. Infatti, la stagionalità è per definizione un'oscillazione ciclica e quindi, una volta che è stata determinata, essa resterà costante nel tempo per un determinato sensore. D'ora in avanti la nostra attenzione si concentrerà quindi sul residuo (chiamato *Diff* – "Differenziale" nel codice), considerando la stagionalità nota per ogni sensore. Indubbiamente su intervalli di tempo molto lunghi anche la stagionalità può evolvere, ma potrà sempre essere corretta periodicamente a partire dalle serie storiche e non necessiterà in ogni caso una previsione in tempo reale.

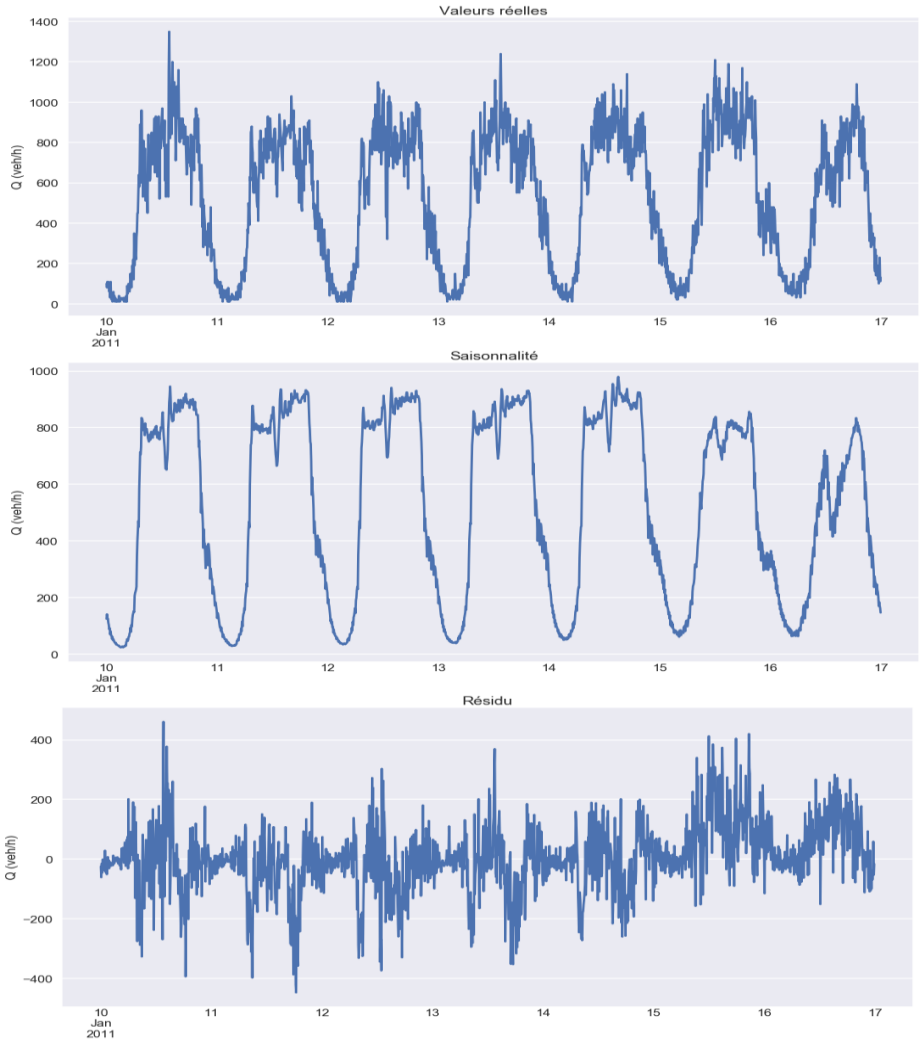


Figura 13: Esempio della disaggregazione della serie storica su una settimana del sensore 226 di Lione

Queste riflessioni sulla disaggregazione in stagionalità, trend e residuo valgono soprattutto per quanto riguarda i valori di portata, mentre il comportamento meno regolare del tasso di occupazione (valori normalmente intorno al 10% con picchi isolati fino oltre il 60%) fa in modo che la disaggregazione non comporti vantaggi significativi.

b. Variabili dipendenti dal calendario

Il modello necessita di variabili che permettano di affinare la previsione secondo l'ora della giornata, il giorno della settimana, il mese e tutti gli altri effetti legati al calendario, sia che si voglia prevedere il residuo della portata o il tasso di occupazione.

Queste variabili vengono create prima della fase di adattamento del modello attraverso un trattamento dei dati di input. La disponibilità di un *timestamp*, che colloca temporalmente ogni

misura, agevola la creazione sia di variabili booleane che di variabili numeriche. Le variabili scelte per questo progetto sono:

- "Mois"; variabile numerica da 1 a 12 che identifica il mese
- "Jour": variabile numerica da 1 a 7 che identifica il giorno sella settimana
- "Heure": variabile numerica da 1 a 240 che identifica la fascia di 6 minuti della giornata
- "Heure de pointe": variabile booleana che identifica l'istante considerato come appartenente o meno all'ora di punta di quel sensore. L'ora di punta è un parametro difficile da determinare: per lo stesso sensore possono esserci una punta al mattino e una la sera, nel week-end l'ora di punta può non avere significato, ecc. Nel nostro caso l'ora di punta è stata individuata in maniera molto semplificata scegliendo l'istante in cui si è registrato il valore massimo di portata nella giornata e considerando i 30 minuti precedenti e i 30 minuti seguenti. Dato l'elevato grado di semplificazione nella definizione della variabile, questa si è rivelata ininfluente per il modello ed è quindi stata accantonata. Ciononostante, un approfondimento che porti ad una definizione più solida e pertinente di questa variabile potrà essere oggetto di future migliorie del modello.
- "Fêtes": variabile booleana che indica se il giorno considerato è festivo o feriale
- "Vacances": variabile booleana che indica se il giorno considerato è un giorno di vacanza scolastica.

c. Portata e tasso di occupazione agli istanti precedenti

Una volta definite tutte le variabili che prendono in considerazione gli effetti del calendario, i valori più importanti su cui si basa la previsione sono le misure in tempo reale. A partire dalle condizioni di traffico osservate ad un istante t, l'obiettivo diventa prevederne il valore all'istante $t+n$, che indica l'orizzonte di previsione scelto. Per questo progetto l'obiettivo era una previsione a 30 minuti, dunque l'algoritmo può avvalersi al massimo delle misure a $t-30$ minuti, ovvero $t-5$ intervalli di 6 minuti.

Per una previsione dell'istante t, al modello sono forniti i valori dell'ultima ora (da $t-5$ a $t-15$) e i valori della mezz'ora da prevedere misurati il giorno precedente (da $t-240$ a $t-245$). All'inizio della sperimentazione sono state fornite al modello anche le rilevazioni di 12 ore prima (da $t-120$ a $t-125$) ma il loro contributo alla previsione si è rivelato poco significativo. Per questa ragione e per limitare il numero di variabili rendendo il calcolo più

leggero e veloce, tali misure sono in seguito state tralasciate. L'importanza delle variabili appena descritte nel processo previsionale del modello è illustrata in Figura 14, dove quelle evidenziate in rosso rappresentano proprio le misure che sono state escluse nel seguito della sperimentazione.

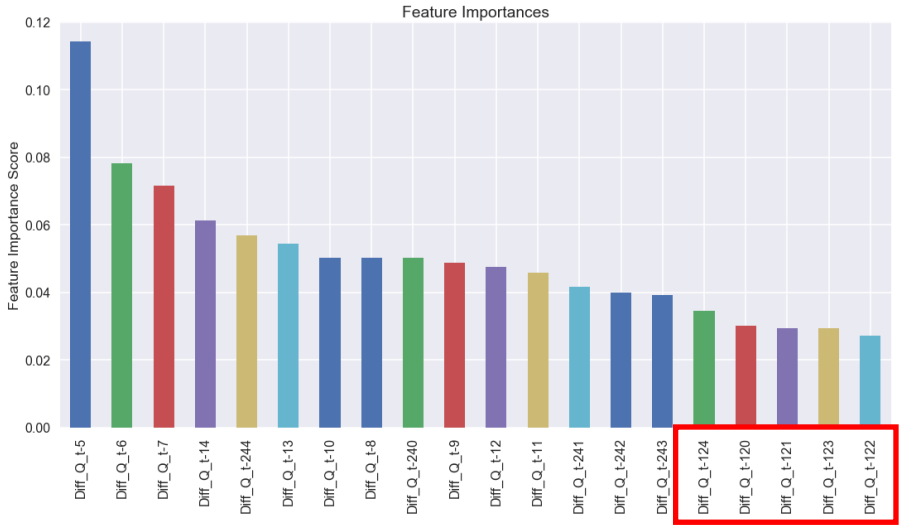


Figura 14: Feature importance delle misure degli istanti precedenti ($t - n$) per la previsione della portata (spire DiRIF)

Il contributo delle singole variabili alla previsione è calcolato tramite il modulo *feature_importance*. Tale modulo utilizza un metodo chiamato "*mean decrease impurity*" [24]: il contributo è espresso come un valore compreso tra 0 e 1 ed indica l'importanza di una variabile nel processo decisionale dell'algoritmo.

Rappresentando l'importanza di tutte le variabili temporali sullo stesso grafico si può valutare quali sono i fattori più importanti per la previsione del traffico. I risultati mostrano, in ambito urbano come in ambito extra-urbano, che la variabile più importante è l'ora della giornata. In seguito, i risultati differiscono tra i due casi: in ambito urbano la portata della mezz'ora precedente è molto importante (da $t - 5$ a $t - 9$) mentre le portate più vecchie (da $t - 10$ a $t - 14$) hanno un'importanza comparabile con i valori del giorno precedente (da $t - 240$ a $t - 244$); al contrario per una rete più lineare le informazioni del giorno precedente sono meno interessanti. In entrambi i casi poi le informazioni riguardo al giorno della settimana e al mese hanno ancora rilevanza, mentre le variabili booleane non contribuiscono quasi per nulla alla previsione.

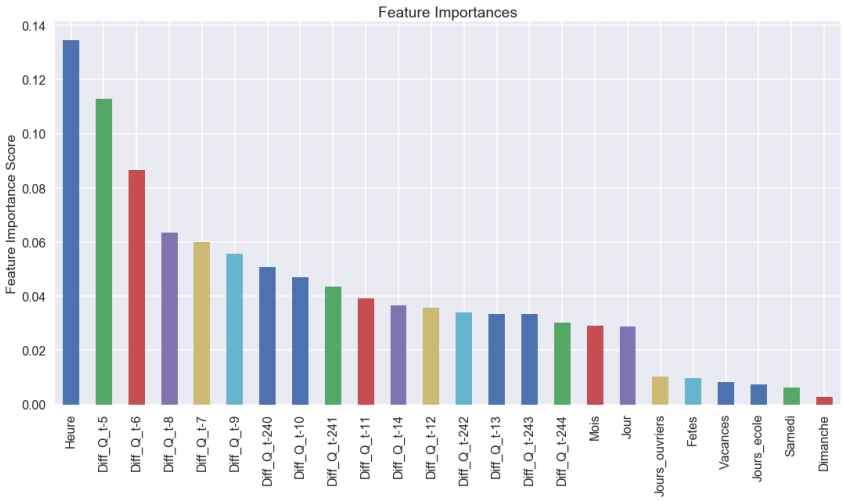


Figura 15: Grafico riassuntivo dell'importanza delle variabili temporali – (dati di Lione)

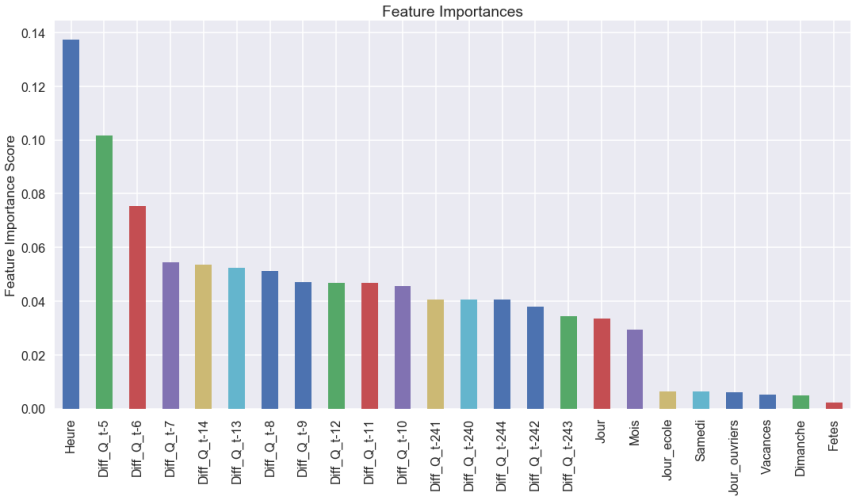


Figura 16 : Grafico riassuntivo dell'importanza delle variabili temporali – (dati della DiRIF)

2. Aspetti spaziali

Dopo aver analizzato gli aspetti temporali passiamo all'analisi degli aspetti spaziali. Il traffico su un segmento stradale ad un determinato istante dipende fortemente dal traffico sui segmenti adiacenti negli istanti precedenti. Per una buona previsione della congestione l'algoritmo dovrà quindi appoggiarsi anche sulle rilevazioni di portata (o di tasso di occupazione) degli altri sensori della rete. Non è tuttavia possibile fornire al modello le misure di tutti i sensori della

rete, in quanto i tempi di calcolo diventerebbero troppo lunghi per un utilizzo in tempo reale dello strumento. Il problema diventa quindi quali sensori considerare, ovvero individuare i segmenti che influenzano il traffico in un punto specifico della rete.

Per la rete extraurbana dell'Ile-de France il problema è di più semplice risoluzione in quanto, trattandosi prevalentemente di assi lineari, è logico considerare i valori misurati su tutto il segmento a monte e a valle del punto di interesse, o almeno su una sua parte.

Per una rete urbana, come quella di Lione, è ben più complesso definire quali segmenti sono più importanti per la previsione in un dato punto. Per superare questa difficoltà si è fatto ricorso a metodi di apprendimento non supervisionato o clustering. [25]

a. Clustering per la rete urbana di Lione

Il clustering è una tecnica di classificazione che permette di raggruppare gli individui in clusters. Un "*cluster*" è definito come una classe di oggetti che sono simili tra loro e diversi da quelli delle altre classi. Gli algoritmi di clustering sono utilizzati per esempio in biologia per derivare le tassonomie degli animali e delle piante o nel marketing per identificare dei gruppi di consumatori. L'algoritmo valuta la similarità tra gli individui attraverso la loro prossimità una volta che le loro caratteristiche sono rappresentate in uno spazio multidimensionale. Tale prossimità o vicinanza è espressa tramite una distanza euclidea calcolata proprio sui parametri forniti all'algoritmo. In letteratura esistono numerose tecniche di clustering: la principale differenza tra tali tecniche risiede nell'algoritmo utilizzato nella definizione delle classi (*K-means*, *Affinity propagation*, *Agglomerative clustering*, ecc.). Nel caso di questo studio si è scelto di utilizzare un algoritmo di tipo *K-means,* anche detto algoritmo dei K vicini più prossimi. L'applicazione dell'algoritmo prevede:

- la definizione del numero M di cluster da ottenere e del parametro K;
- la costruzione di una matrice che ha per linee tutti i punti della rete e per colonne tutti gli istanti per cui è disponibile una rilevazione. Le misure di portata (o del tasso di occupazione) sono dunque gli elementi della matrice. Su tale matrice ha luogo il calcolo dei vicini più prossimi: i valori della portata saranno confrontati tra loro per ritrovare delle caratteristiche comuni;
- la stima del peso dei vicini più prossimi sulla base delle distanze euclidee;

– la divisione dei punti in M clusters.

Il risultato del clustering è una divisione della città di Lione in 20 "quartieri" che hanno, secondo l'algoritmo, delle caratteristiche di traffico in comune. Poiché la classificazione viene realizzata in automatico basandosi esclusivamente su calcoli statistici, sta all'analista interpretare i risultati e eventualmente procedere per tentativi variando i valori dei parametri M e K fino a che la classificazione non appare verosimile.

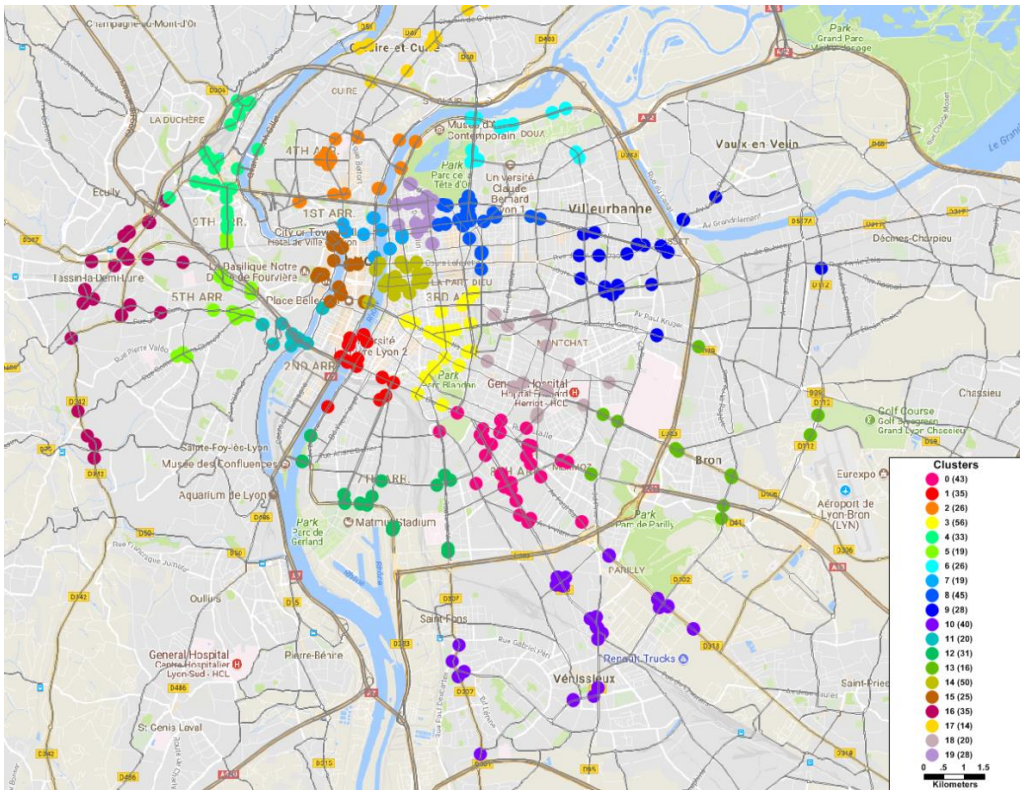


Figura 17: Divisione della rete di sensori di Lione in 20 clusters

Per prevedere le caratteristiche del traffico di un punto della rete urbana di Lione saranno quindi utilizzate tutte le misure dei sensori appartenenti a quel determinato cluster.

In ambito urbano come in ambito extra-urbano, le misure dei sensori adiacenti a quello di interesse sono state fornite solo per l'istante $t - 5$. La feature importance mostra che tale variabile ha un impatto non trascurabile sulla previsione, o addirittura molto importante nel caso della rete DiRIF (cf. Figura 18, Figura 19).

3. Ulteriori aspetti

Le ultime variabili considerate nella modellizzazione sono quelle che descrivono i fattori esterni al traffico. Tali fattori sono spesso impossibili da prevedere, come nel caso degli incidenti. C'è tuttavia una categoria che può essere prevista con estrema precisione: gli eventi meteorologici.

I dati meteo sono stati quindi introdotti nel modello, utilizzando quelli disponibili in open data sul sito di Meteo France [26]. I dati sono disponibili per fasce di tre ore e contengono un volume di informazioni ridondante. Le variabili ricavate da questi dati sono:

- "Pluie": una variabile booleana che indica se la precipitazione piovosa attesa nelle tre ore successive è più intensa di 2 mm. Il valore di 2 mm è stato scelto come discriminante tra una precipitazione rilevante o non rilevante.

- "Neige": una variabile booleana che indica se è attesa una precipitazione nevosa nelle tre ore successive

I risultati ottenuti mostrano che l'impatto di queste due variabili sul modello è trascurabile: il problema principale è probabilmente la durata della finestra della previsione che, come detto, è di tre ore, troppo grande rispetto alle fasce di sei minuti dei dati di traffico. In ogni caso un'analisi più approfondita dell'impatto delle rilevazioni meteorologiche potrà costituire una pista di miglioramento dello strumento.

A conclusione del paragrafo relativo alle variabili esplicative, Figura 18 e Figura 19 riassumono l'importanza di tutte le variabili definite finora e richiamate in Tabella 5.

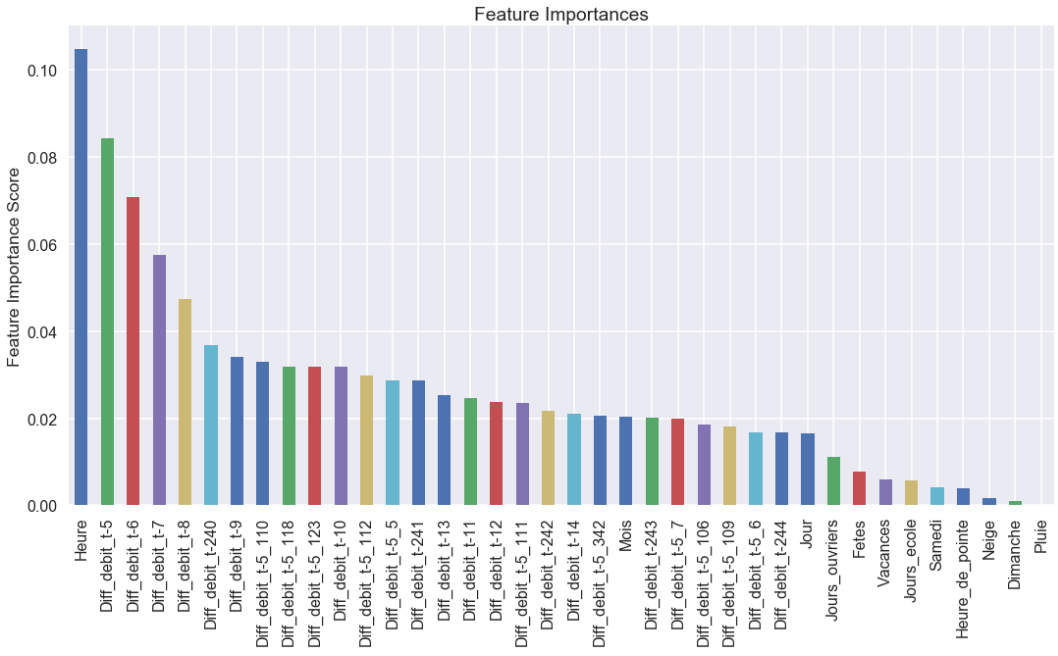


Figura 18: Riassunto dell'importanza di tutte le variabili definite (dati di Lione)

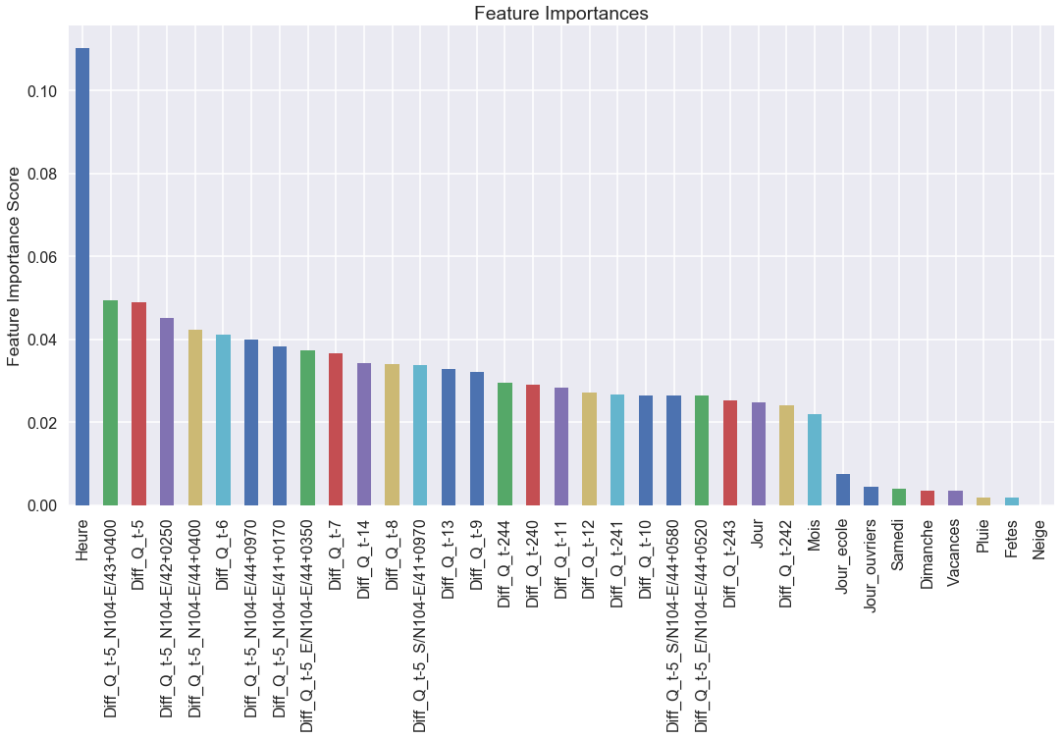


Figura 19: Riassunto dell'importanza di tutte le variabili definite (dati della DiRIF)

Diff_Q_t-n	Numerica	Differenziale della portata all'istante $t-n$ per il sensore considerato
Diff_Q_t-n_X	Numerica	Differenziale della portata all'istante $t-n$ per il sensore X dello stesso cluster / segmento stradale
Heure	Numerica	Fascia oraria da 6 minuti della giornata (valori da 1 a 240)
Jour	Numerica	Giorno della settimana (valori da 1 a 7)
Mois	Numerica	Mese (valori da 1 a 12)
Samedi	Booleana	"True" se il giorno considerato è un sabato
Dimanche	Booleana	"True" se il giorno considerato è una domenica
Vacances	Booleana	"True" se il giorno considerato è un giorno di vacanze scolastiche
Fêtes	Booleana	"True" se il giorno considerato è un giorno festivo
Jours_ouvrables	Booleana	"True" se il giorno considerato non è né sabato, né domenica, né festivo
Jours_ecole	Booleana	"True" se il giorno considerato non è né un sabato, né una domenica, né un giorno di vacanze scolastiche
Pluie	Booleana	"True" se è pervista una precipitazione piovosa di più di 2 mm
Neige	Booleana	"True" se è prevista una precipitazione nevosa

Tabella 5: Tabella riassuntiva delle variabili esplicative del modello

C. Previsione della portata e del tasso di occupazione

Per la previsione della portata e del tasso di occupazione a 30 minuti sono stati utilizzati degli algoritmi di apprendimento supervisionato. In Python esistono diverse librerie che mettono a disposizione numerosi algoritmi di questo tipo. [27] Alcuni di questi algoritmi sono stati applicati al modello per confrontarne le prestazioni. I risultati che hanno condotto alla scelta di un algoritmo in particolare, saranno presentati nel paragrafo V.A.

Indipendentemente dall'algoritmo utilizzato, il metodo di applicazione è sempre lo stesso. Le principali tappe della sua applicazione sono descritte nel seguito.

1. Indicatori di performance

È necessario innanzitutto definire i criteri di performance che saranno utilizzati per confrontare e valutare i risultati. Gli indicatori sono:

- **r^2 o coefficiente di determinazione.** Nel caso di una regressione lineare è espresso come:

$$r^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{n-1}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n-1}(y_i - \bar{y}_i)^2}$$

dove n è il numero di misure, y_i è il valore reale di una misura, $\bar{y}_i$ il valor medio delle misure reali e $\hat{y}_i$ il valore previsto dall'algoritmo. [29] Tale coefficiente può essere definito come il rapporto tra la varianza esplicata e la varianza totale. Il suo valore è compreso tra 0 e 1: un valore di 1 significa che tutta la varianza è stata esplicata dal modello, e che quindi le sue prestazioni sono ottime.

- **MSE (*mean squared error*) o errore quadratico medio.** Si esprime come:

$$MSE\ (y, \hat{y}) = \frac{1}{n}\sum_{i=0}^{n-1}(y_i - \hat{y}_i)^2$$

dove n è il numero di misure, y_i è il valore reale di una misura e $\hat{y}_i$ il valore previsto dall'algoritmo. [27] In realtà il parametro utilizzato è l'RMSE, ovvero l'MSE sotto radice quadrata. Tale operazione permette di ottenere degli errori che hanno lo stesso ordine di grandezza delle misure reali. Questo parametro non permette di prendere in considerazione la direzione dell'errore, ma solo la sua ampiezza.

- **MAE (*mean absolute error*) o errore assoluto medio.** Si esprime come:

$$MAE\ (y, \hat{y}) = \frac{1}{n}\sum_{i=0}^{n-1}|y_i - \hat{y}_i|$$

dove n è il numero delle misure, y_i è il valore reale di una misura e $\hat{y}_i$ è il valore previsto dall'algoritmo. [27] Come nel caso precedente anche questo parametro non permette di prendere in considerazione la direzione dell'errore ma solo la sua ampiezza a causa della presenza del valore assoluto. Rispetto all'MAE, l'RMSE è più sensibile ai grandi errori che vengono elevati al quadrato.

2. Iperparametri

Dopo aver definito i criteri di performance, sono stati testati diversi tipi di algoritmi e quello che ha fornito i risultati migliori è stato scelto per il prosieguo del progetto. Per migliorare le prestazioni di tale algoritmo, lo si è sottoposto ad un processo di *grid search*, ovvero di definizione degli iperparametri. Quest'ultimi sono definiti come dei parametri esterni, tipici di un modello, che intervengono per garantire una buona generalizzazione mantenendo al tempo stesso la precisione. Concretamente, il grid search consiste nell'applicazione dello stesso algoritmo più volte, cambiando il valore di un iperparametro per volta. Ciò permette la scelta del miglior valore per ogni iperparametro, o più precisamente la loro migliore combinazione. Una scelta infelice degli iperparametri può portare alla comparsa di problemi di overfitting e di underfitting.

- l'*overfitting* (sovra-apprendimento) si verifica quando un modello è troppo calzante ai dati su cui si è allenato. Si traduce nell'incapacità di fornire delle buone previsioni quando si trova di fronte a nuovi dati sui quali non si è allenato.

- l'*underfitting* (sotto-apprendimento), al contrario, si verifica quando un modello è troppo generalizzato. Si traduce in errori importanti su tutte le previsioni.

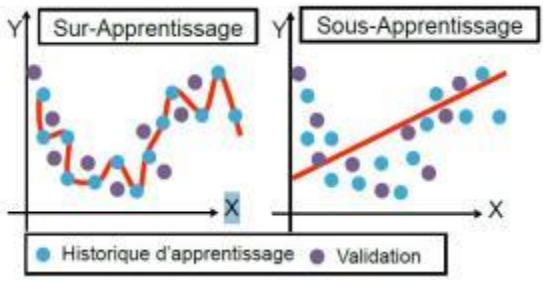


Figura 20: Rappresentazione grafica dell'overfitting e dell'underfitting

Fonte: [2]

3. Cross-validation

La *cross-validation* o validazione incrociata è una tecnica che permette di testare lo stesso algoritmo più volte e confrontarne i risultati senza incorrere nell'overfitting avendo a disposizione una sola base di dati.

Per poter valutare le prestazioni di un algoritmo predittivo è innanzitutto necessario un adattamento su una serie temporale nota. Ma la previsione successiva all'adattamento non può essere fatta sugli stessi dati, perché l'algoritmo li conosce già essendocisi "allenato" in precedenza. La conseguenza sarebbe una sovrastima delle sue prestazioni. In più si forzerebbe un overfitting dall'esterno, rendendo impossibile l'individuazione di questo problema. Per ovviare a ciò, la base di dati (*data-set*) viene divisa in due parti: un set di allenamento (*train-set*) e un set di validazione (*test-set*). Le dimensioni dei due set sono generalmente in proporzione 80-20 %, ma se la base di dati è molto grande le proporzioni possono passare a 90-10 %.

La cross-validation interviene durante il confronto dei diversi algoritmi o durante il grid search e assicura che gli indicatori di performance scelti siano affidabili. Tali indicatori sono calcolati come media su diversi tentativi: ma con un solo data-set si può applicare l'algoritmo una sola volta. La cross-validation consiste quindi nel dividere il data-set in tante parti e utilizzare queste parti alternativamente come train-set e come test-set. Ciò permette di applicare lo stesso algoritmo più volte e di fare successivamente una media degli indicatori di performance, il tutto conservando un unico data-set.

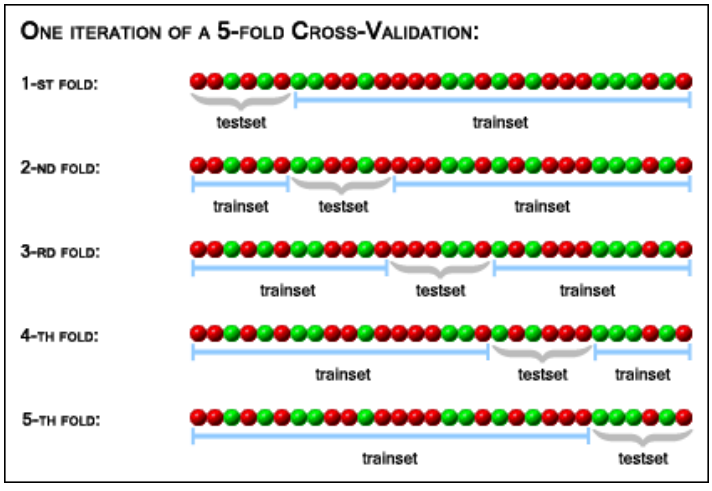


Figura 21: Rappresentazione schematica della cross-validation

Fonte: ProClassify User's Guide

D. Definizione della congestione stradale

L'ultima tappa è la definizione di una variabile che permetta di identificare la congestione. L'obiettivo è trovare una definizione della congestione che sia valevole per tutti i punti della rete e che permetta così di automatizzare la distinzione tra traffico congestionato e non congestionato. Tale definizione deve necessariamente basarsi sui valori di portata e di tasso di occupazione che sono stati previsti ad ogni istante per tutti i punti.

Un'analisi preliminare è stata condotta sui dati reali della DiRIF, in generale di migliore qualità rispetto a quelli di Lione, per tentare di osservare visualmente la congestione sul diagramma fondamentale. L'analisi è stata fatta su dei diagrammi portata-velocità perché questi sono di più facile e immediata lettura, ma gli stessi risultati si possono osservare su un diagramma portata-densità che corrisponde alle variabili disponibili (ricordandosi che la densità è direttamente proporzionale al tasso di occupazione). I risultati illustrati in Figura 22 mostrano l'evoluzione del traffico durante una giornata tipo (giorno feriale) dalle 6:00 alle 21:00 per uno specifico sensore.

Per questo particolare sensore l'ora più congestionata va dalle 8:00 alle 9:00, quando le velocità sono basse e le portate elevate. Ci troviamo nella parte bassa della curva, al di sotto della velocità critica. Al contrario, tra le 6:00 e le 7:00 il traffico è poco congestionato: le velocità sono elevate e la densità è bassa.

Applicando un algoritmo di clustering sulle nuvole di punti corrispondenti del diagramma portata-tasso di occupazione, è possibile dividere tale diagramma in due parti (Figura 23). La prima parte (in blu) è formata dai punti che appartengono al cluster 6:00-7:00 e che rappresentano quindi uno stato non congestionato, mentre la seconda parte (in rosso) è formata da punti che appartengono al cluster 8:00-9:00, ovvero uno stato congestionato.

Con i dati ottenuti dalla previsione effettuata in precedenza, è stata ricostruita una curva portata-tasso di occupazione. In seguito su di essa è stato applicato il clustering, che ha fornito la divisione tra stato congestionato e non congestionato (Figura 24).

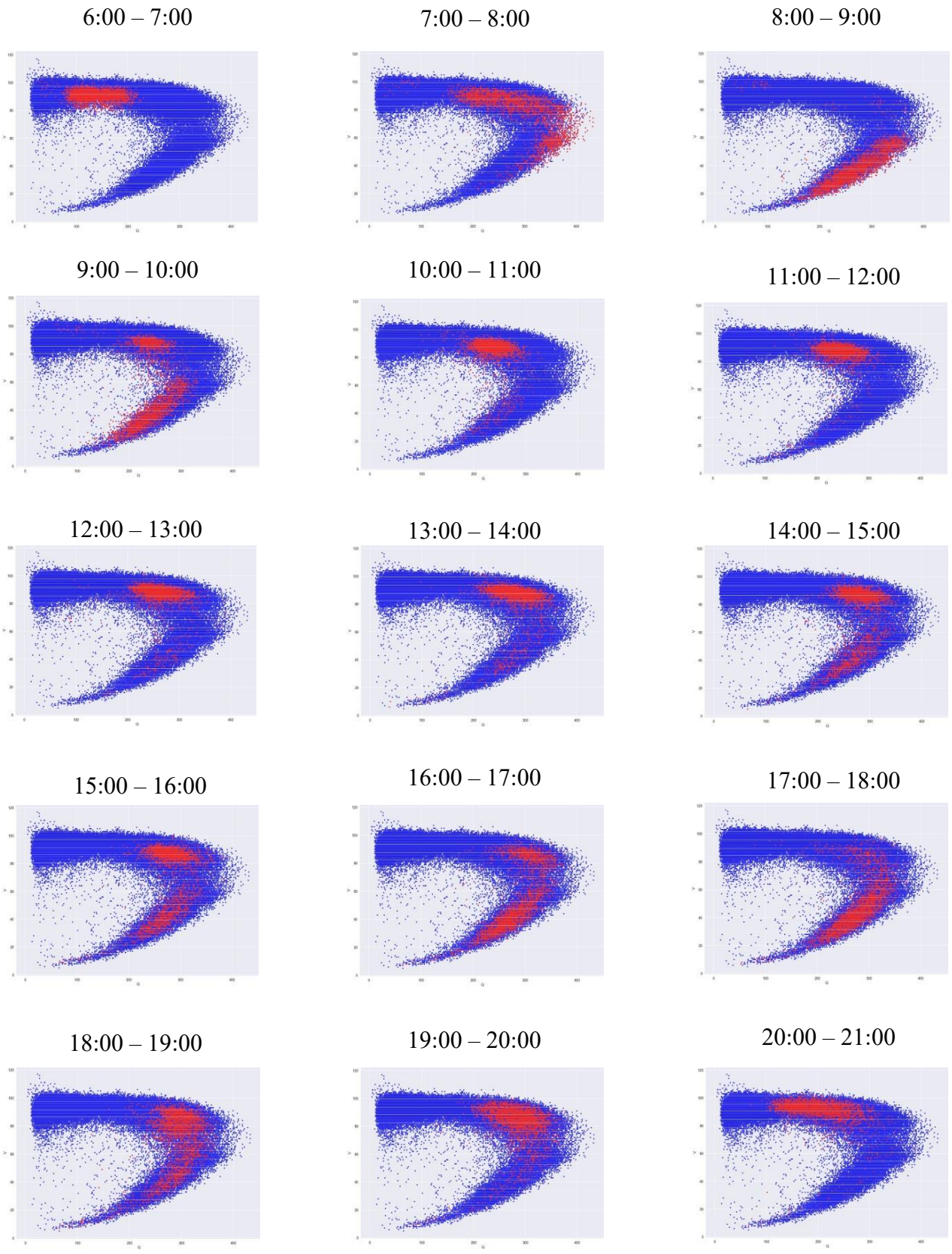


Figura 22: Evoluzione della congestione dalle 6:00 alle 21:00 di un giorno feriale

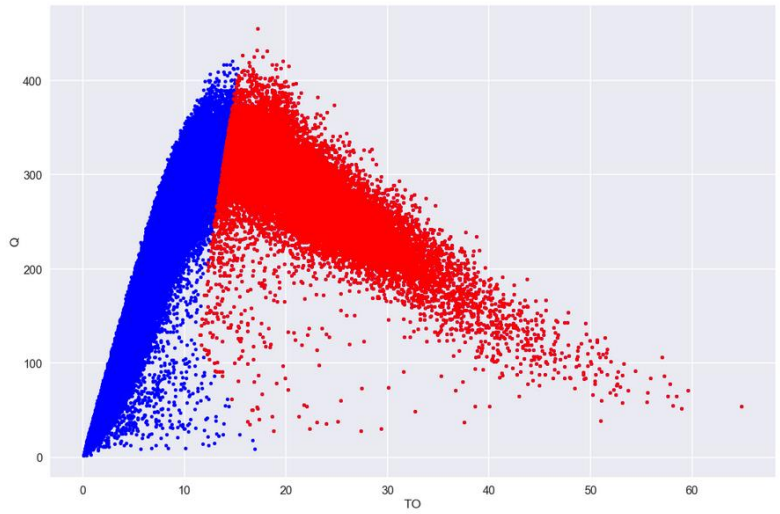


Figura 23: Curva dei valori osservati, in evidenza lo stato congestionato (in rosso) e non congestionato (in blu)

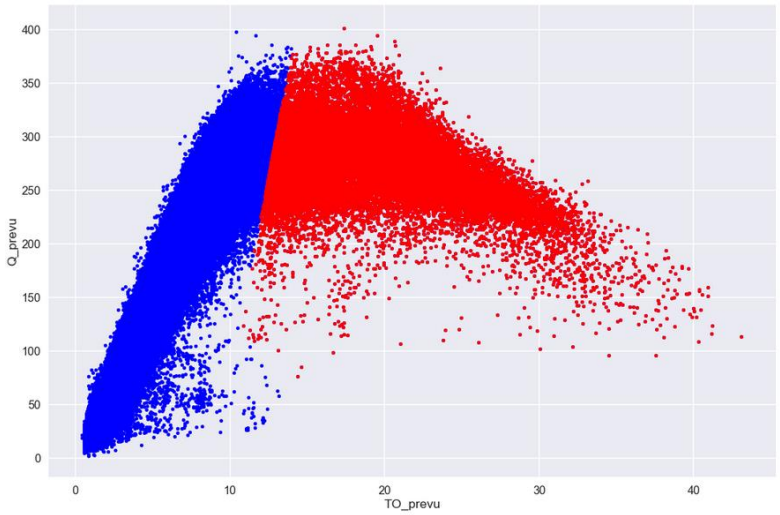


Figura 24: Curva dei risultati della previsione, in evidenza lo stato congestionato (in rosso) e non congestionato (in blu)

V. Risultati

L'obiettivo di questo capitolo è di presentare i risultati ottenuti durante la sperimentazione, in particolare quelli riguardanti la scelta dell'algoritmo, la previsione della portata e del tasso di occupazione e l'identificazione degli stati di congestione. Inoltre, verranno individuati i punti deboli del modello e saranno presentate le future azioni di miglioramento che permetteranno di intervenire per ottimizzarlo.

A. Scelta dell'algoritmo di regressione

La principale famiglia di algoritmi testati è quella degli algoritmi di regressione, in particolare:

- un algoritmo di regressione lineare (*LinearRegressor*, chiamato nel seguito LR): si tratta di una semplice regressione lineare in cui il modello cerca di minimizzare la somma residuale dei quadrati tra i valori osservati e i valori previsti

- un algoritmo di regressione tramite un albero di decisione (*DecisionTreeRegressor*, chiamato nel seguito DTR): si tratta di una regressione fatta a partire da una cascata di test binari (i rami dell'albero) sulle variabili che descrivono il fenomeno (le foglie). In generale ad ogni nodo interno all'albero, si ha un test binario su una variabile tra quelle presenti nella base e due rami che corrispondono alle due possibili soluzioni del test. I test all'interno di un albero decisionale sono fatti a partire dalle informazioni ottenute durante la fase di adattamento del modello. Il cammino dalla radice ad un nodo finale dell'albero corrisponde a una serie di test e alle loro risposte. [25]

- un algoritmo di regressione a foresta di decisioni (*RandomForestRegressot*, chiamato in seguito RFR): si tratta dell'evoluzione del modello precedente. Al posto di un solo albero decisionale, per ogni valore da prevedere sono valutati più alberi (è il principio di una foresta composta da più alberi) e il risultato è la media dei risultati di ogni albero. Il vantaggio di questo algoritmo è la sua maggiore affidabilità che riduce il rischio di overfitting. Per contro l'algoritmo è più costoso in termini di tempo di calcolo.

- un algoritmo a foresta di decisioni "boosted" (*GradientBoositngRegressor*, chiamato nel seguito GBR): si tratta sempre di un modello a foreste decisionali, ma l'algoritmo è applicato più volte per ottenere una maggiore precisione. Il "boosting" è una ponderazione dell'importanza di ogni valore secondo la precisione con cui è previsto. I valori previsti con minore precisione avranno un peso maggiore nell'applicazione successiva dell'algoritmo. Questo processo è in grado di ridurre ancora di più l'overfitting, a discapito però di un aumento dei tempi di calcolo. [27]

Questi algoritmi sono stati testati e confrontati prima di essere sottoposti al grid search. I risultati presentati nella Tabella 6 rappresentano una media delle previsioni su più punti della rete. Tali risultati si riferiscono alla previsione del solo residuo, in quanto è su questo valore che si può valutare l'effettiva performance dell'algoritmo. La tabella mostra come, in tutti i casi e per tutti gli indicatori, le prestazioni migliori siano quelle del *GradientBoostingRegressor* (GBR). Questo algoritmo ha un r^2 più elevato, ovvero riesce a esplicare una parte maggiore di varianza, e degli errori più piccoli. Il *DecisionTreeRegressor* (DTR) e il *RandomForestRegressor* (RFR) hanno delle prestazioni meno buone del *LinearRegressor* (LR) nonostante siano più sofisticati.

I risultati sul train-set non sono rappresentativi della performance del modello (si tratta di un adattamento e non della previsione), ma permettono di identificare l'overfitting, Una grande differenza tra lo stesso indicatore calcolato sul train-set e sul test-set, come nel caso del DTR e del RFR, indica che il modello è troppo ben calato sui dati del primo e non riesce a essere affidabile sui dati del secondo (su cui è fatta la vera previsione).

Il *GradientBoostingRegressor* è stato scelto come algoritmo per il prosieguo del progetto ed è stato oggetto di un intenso grid search per migliorarne le prestazioni attraverso l'ottimizzazione degli iperparametri.

		R^2 test	RSME test	MAE test	R^2 train	RSME train	MAE train
Portata (diffferenziale)	LR	40,579	85,175	60,785	47,326	90,779	64,148
	DTR	18,043	110,456	80,316	97,556	9,201	5,559
	RFR	24,362	93,450	67,058	92,966	29,275	20,214
	GBR	44,356	82,293	59,293	55,777	87,784	62,186
Tasso di occupazione	LR	50,454	3,656	2,168	65,112	3,900	1,877
	DTR	9,782	4,186	1,923	97,444	0,926	0,372
	RFR	32,770	4,298	2,020	96,624	0,819	0,318
	GBR	59,554	3,260	1,592	70,437	3,595	1,700

Tabella 6: Confronto degli indicatori di performance per i 4 algoritmi testati

B. Applicazione alla rete extra-urbana della DiRIF

Come anticipato nel paragrafo sui dati di input, l'analisi sulla rete DiRIF si è concentrata sulla N104, in particolare sul segmento tra la PR40+0000 e la PR45+0000. Questo segmento della rete si trova sulla parte sud della Francilienne, il terzo il livello di tangenziale che circonda Parigi.

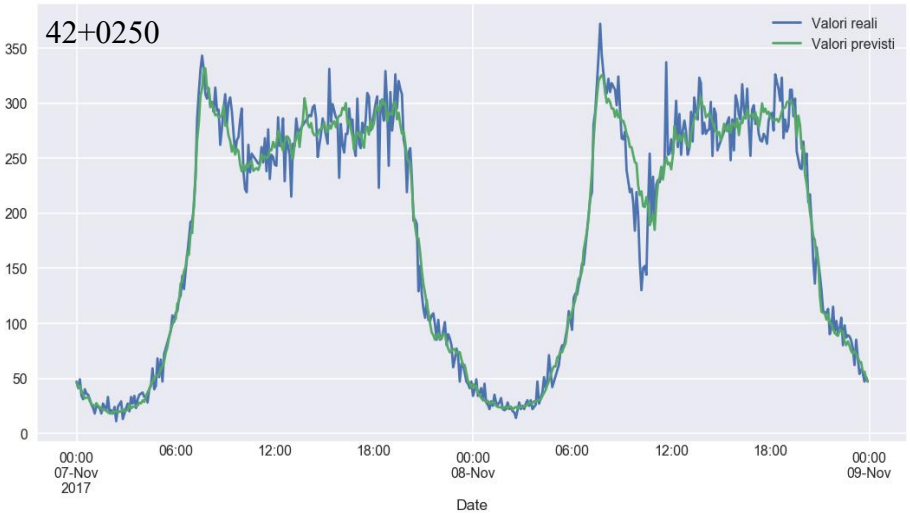
I dati erano disponibili per un periodo di due anni (2016 e 2017): per il test-set si è scelto un periodo di due mesi, novembre e dicembre 2017. I risultati per tre punti rappresentativi di questo tratto stradale (42+025, 43+090, 44+097) sono illustrati nel seguito.

1. Previsione della portata

Le prestazioni del modello per la previsione della portata possono essere calcolate nel caso della portata totale o nel caso più interessante del solo residuo. La qualità dei risultati per la previsione della portata totale è molto buona, con dei valori di r^2 superiori a 95 (Tabella 7). L'andamento della serie prevista è molto simile a quello dei valori misurati anche se, in generale, si osserva la tendenza ad approssimare il valor medio della serie reale (Figura 25 – in blu la serie osservata e in verde la serie prevista). L'algoritmo fornisce quindi una buona stima della portata reale presente sulla rete. L'RMSE e l'MAE mostrano che l'errore medio è inferiore a 20 veicoli sulla fascia di sei minuti (200 veicoli/h), ovvero meno del 10 % della portata totale.

Progressiva	R²	RMSE	MAE
42+0250	95,918	21,118	14,582
43+0900	96,527	19,891	13,704
44+0970	96,811	18,823	12,836

Tabella 7: Indicatori di performance – DiRIF – Portata totale



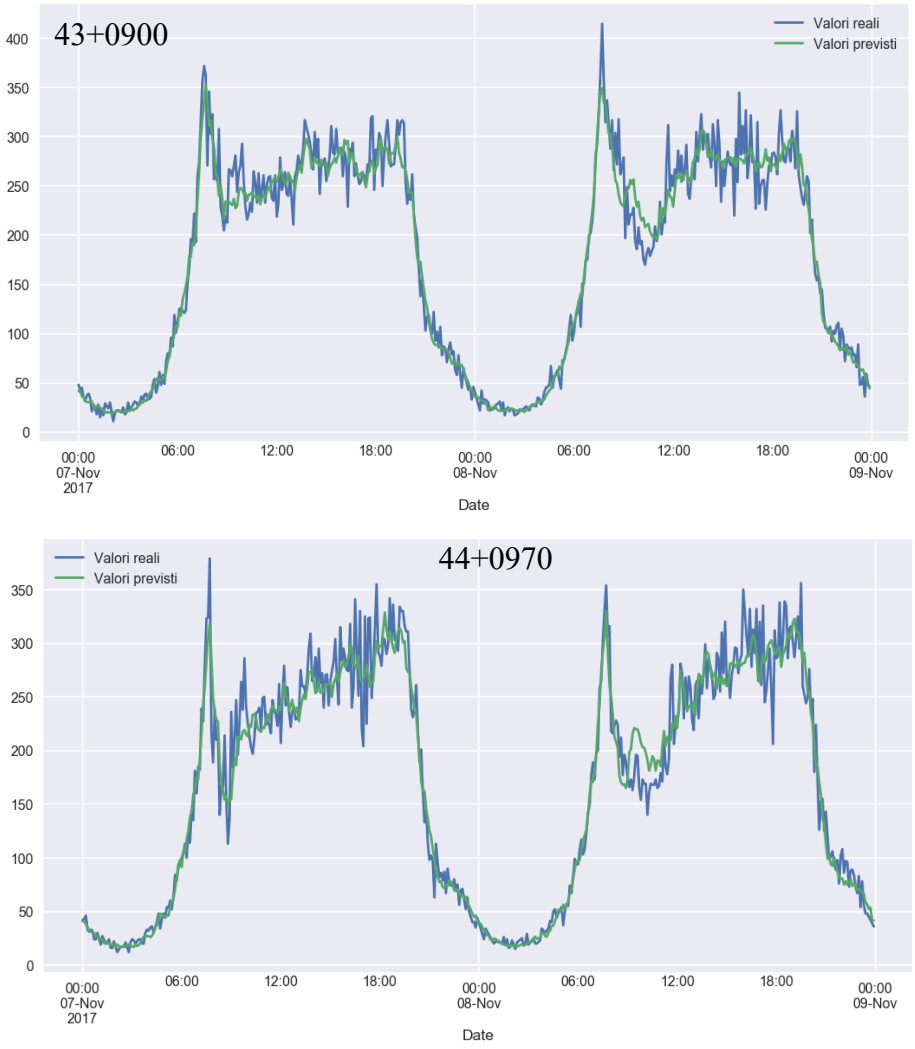
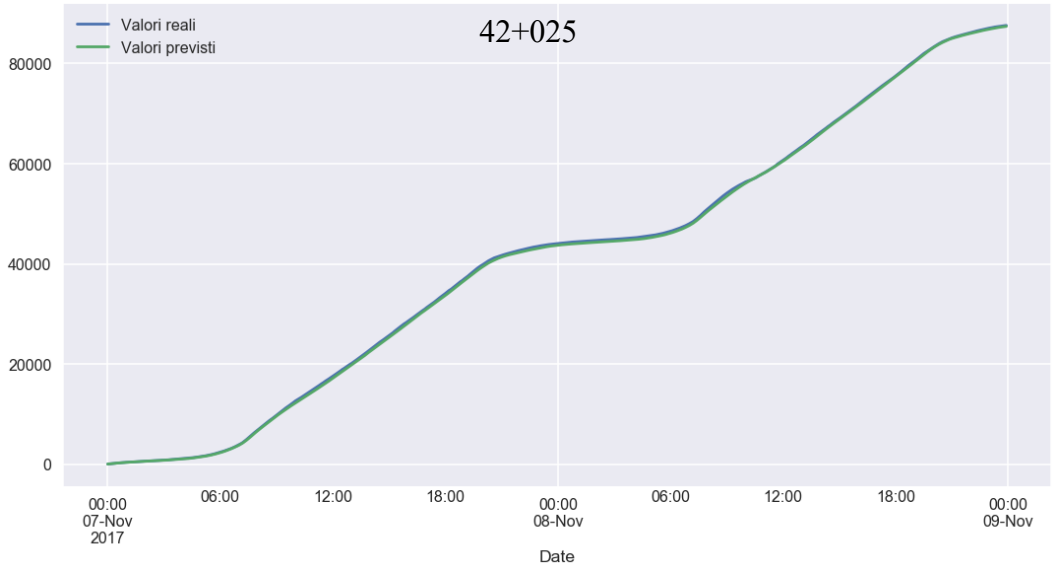


Figura 25: Rappresentazione grafica dei risultati per due giorni di novembre 2017 – DiRIF – Portata totale (veh/6min)



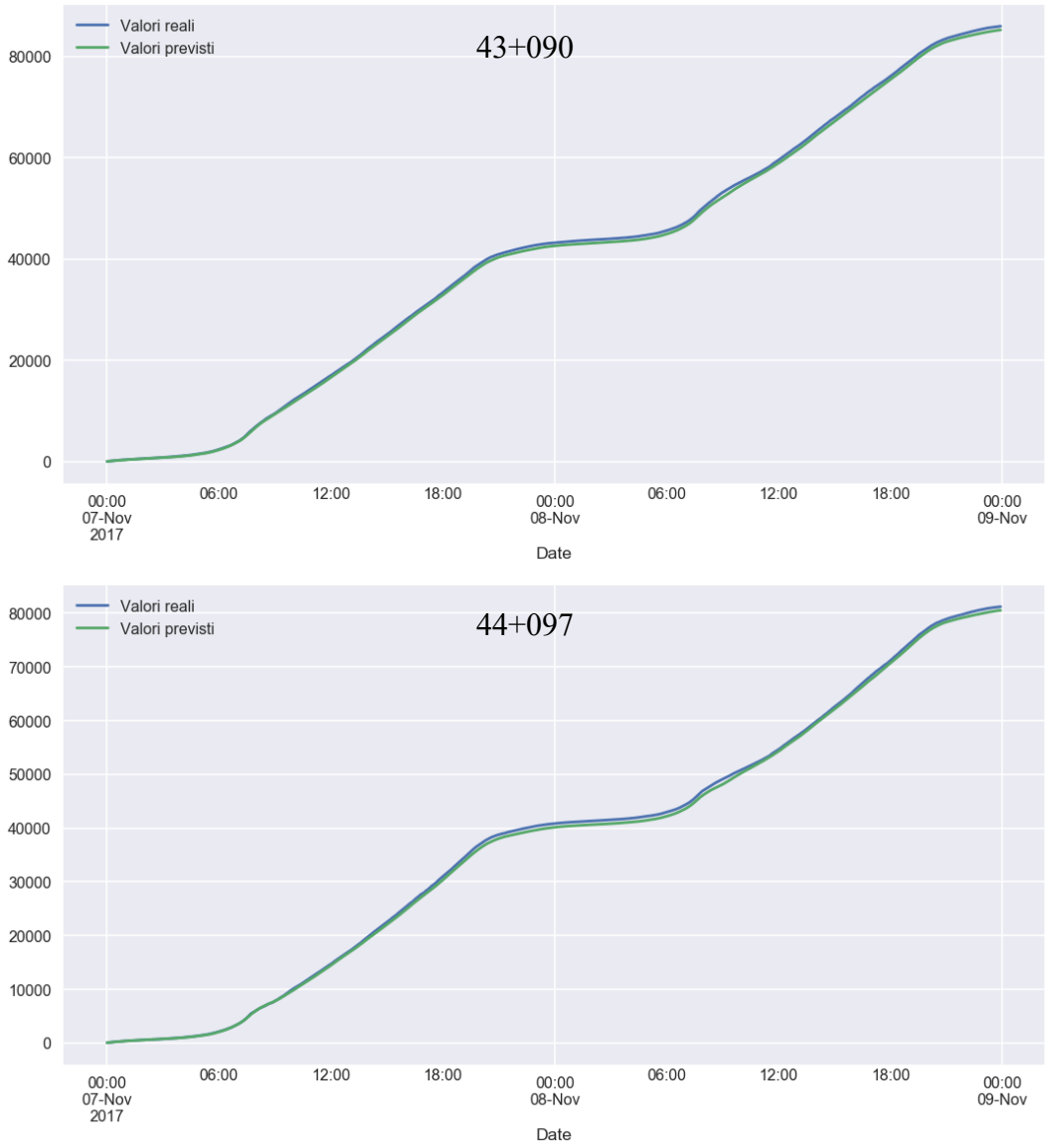
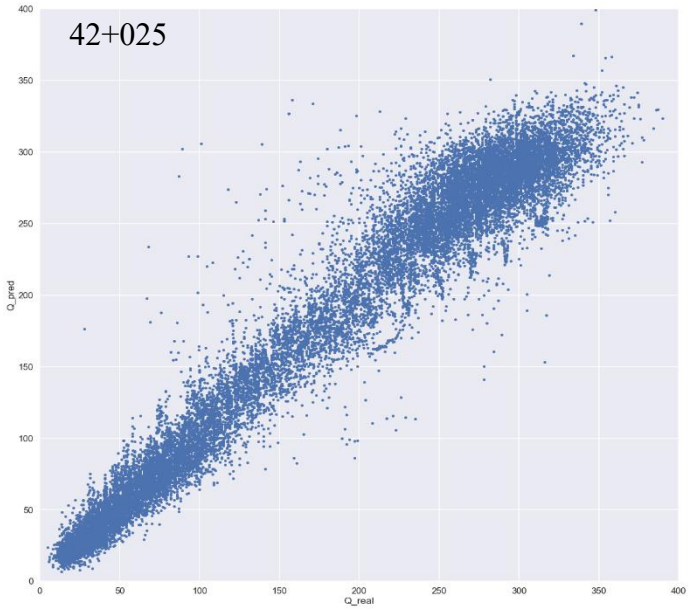


Figura 26: Rappresentazione dei risultati in forma cumulata – DiRIF – Portata totale (veh/6min)



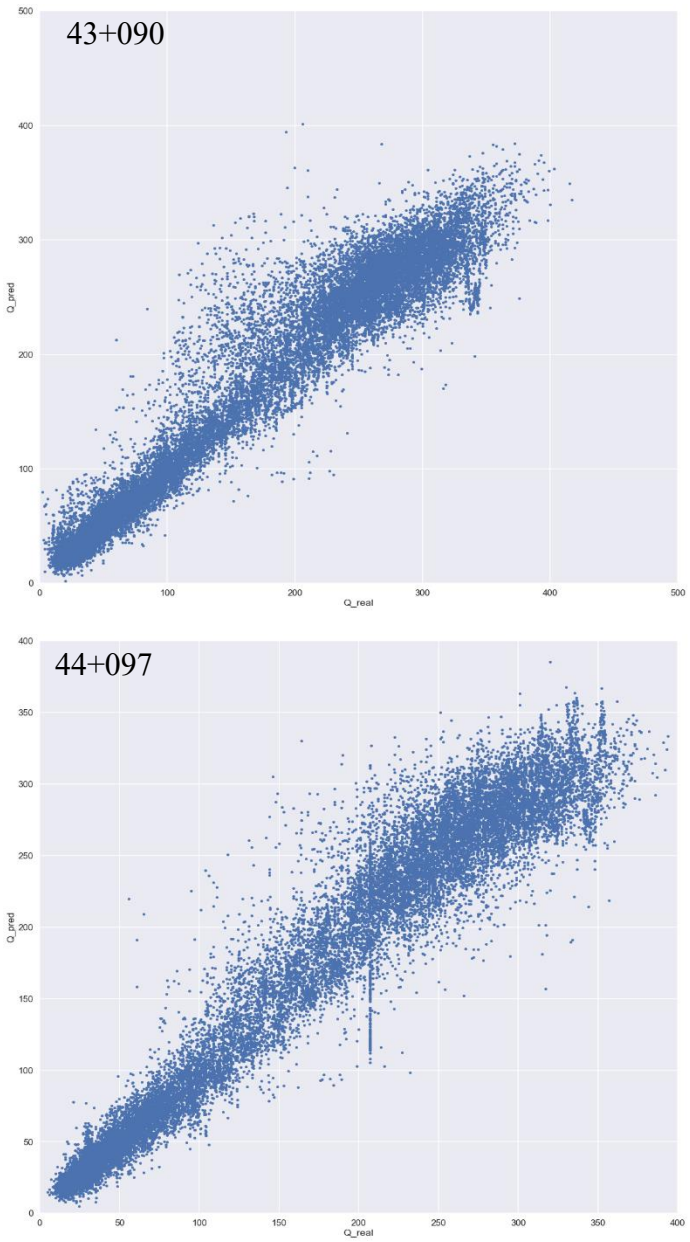


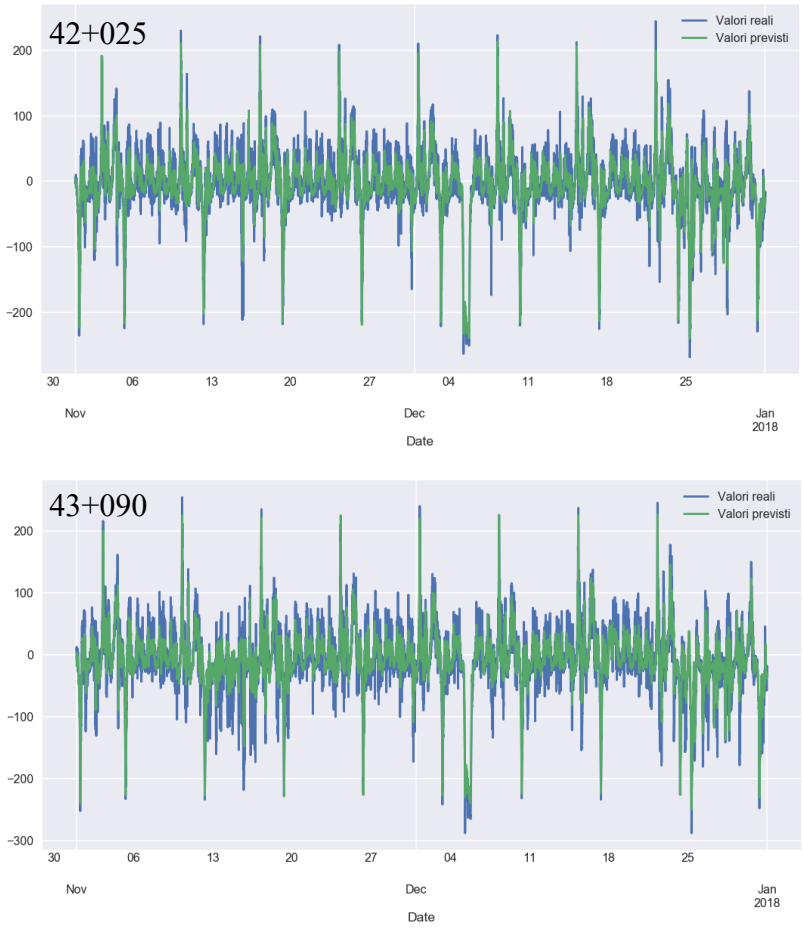
Figura 27: Correlazione tra valori osservati e valori previsti – DiRIF – Portata totale

Analizzando le portate cumulate in Figura 26 ci si accorge che i valori previsti dal modello, in ambito extra-urbano, non divergono rispetto a quelli realmente osservati. La previsione è quindi generalmente affidabile in tutti i momenti della giornata. I grafici in Figura 27 mostrano una buona correlazione tra le portate osservate e quelle previste, confermando la qualità della previsione fornita dal modello. Ciononostante si può osservare una leggera tendenza a sottostimare le portate quando il loro valore è molto grande.

Concentrandosi sulla previsione del solo residuo ci si accorge che il modello ha la tendenza a sottostimare l'ampiezza della varianza (Figura 28). Ciò giustifica la tendenza osservata in precedenza ad approssimare il valor medio delle oscillazioni. È possibile osservare come certe oscillazioni, anche di grande ampiezza, siano molto ben individuate dal modello, in particolare in corrispondenza delle domeniche e dei giorni festivi, mentre altre non sono nemmeno accennate. Ciò può suggerire che le situazioni dove il modello è preciso corrispondono a degli istanti in cui le variabili del modello sono più calzanti. La precisione generale resta buona, anche se ci si accorge che l'RMSE e l'MAE sono gli stessi del caso precedente nonostante i valori di residuo siano molto più piccoli di quelli della portata totale. L'imprecisione del modello è quindi tutta concentrata nella previsione del residuo.

Progressiva	R^2	RMSE	MAE
42+0250	88,690	20,851	14,218
43+0900	87,506	19,891	13,704
44+0970	90,667	21,776	15,285

Tabella 8: Indicatori di performance – DiRIF – Differenziale della portata



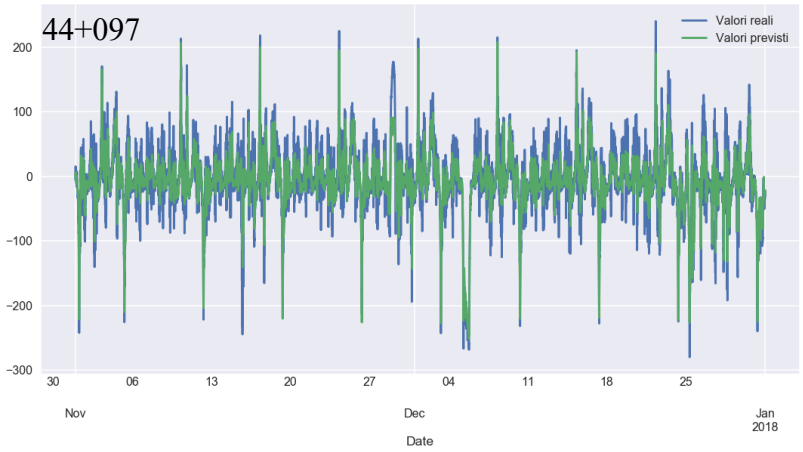


Figura 28: Rappresentazione grafica dei risultati per tutto il test-set – DiRIF – Differenziale della portata (veh/6min)

2. Previsione del tasso di occupazione

I risultati per il tasso di occupazione mostrano un valore di r^2 superiore a 80, quindi molto buono, e degli errori dell'ordine del 3%. Nonostante il tasso di occupazione possa assumere valori compresi tra 0 e 100, la maggior parte dei valori è inferiore al 40%; quindi un errore del 3% significa che per la maggior parte del tempo il modello ha un errore effettivo di poco inferiore al 10%.

Progressiva	R²	RMSE	MAE
42+0250	84,931	3,688	1,690
43+0900	87,850	2,432	1,204
44+0970	79,534	3,505	1,914

Tabella 9: Indicatori di performance – DiRIF – Tasso di occupazione

Osservando il grafico delle serie temporali (Figura 30) ci accorgiamo che il modello ha delle difficoltà nell'identificare l'ampiezza dei picchi più grandi. Ciononostante, data la forma delle curve, sembra più importante che il modello individui la presenza dei picchi e non tanto la loro reale ampiezza: infatti un valore più grande della media (che sia del 30% o dell'60%) identifica già in sé una perturbazione del traffico.

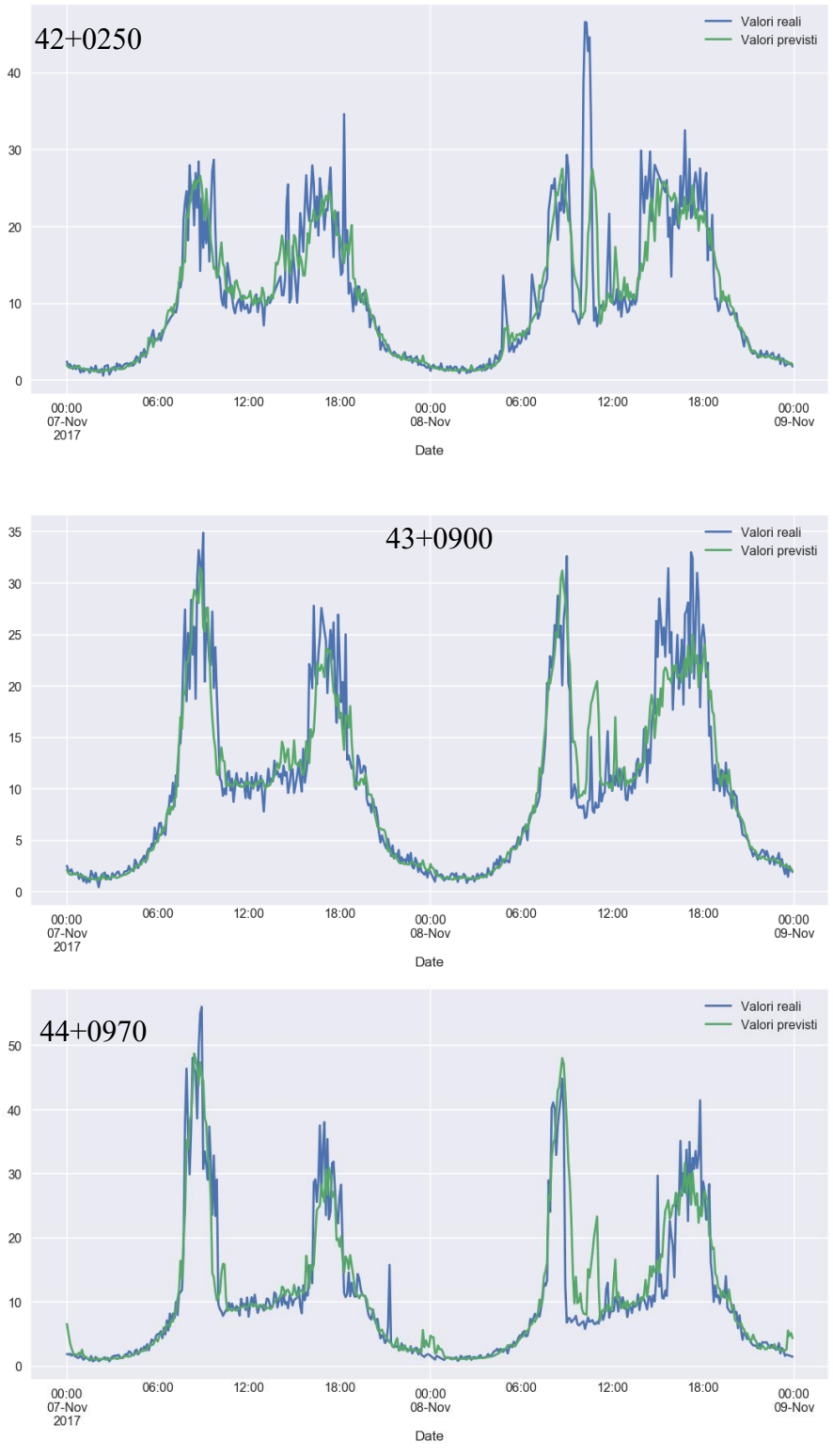


Figura 29 : Rappresentazione grafica dei risultati per due giorni del novembre 2017 – DiRIF – Tasso di occupazione

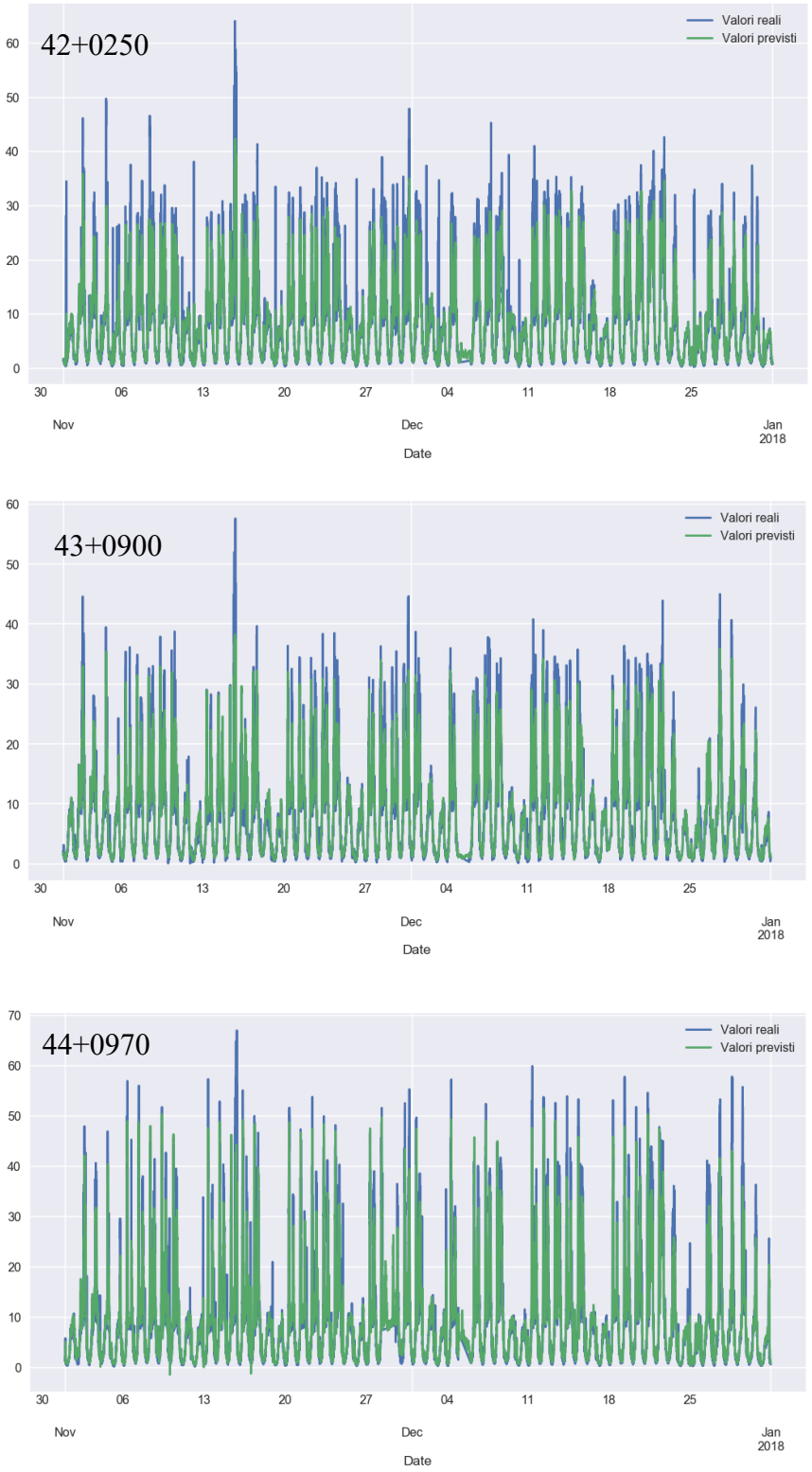


Figura 30: Rappresentazione grafica dei risultati per tutto il test-set – DiRIF – Tasso di occupazione

3. Individuazione della congestione

La performance del modello nell'identificazione dello stato di congestione è valutata, per ogni intervallo di 6 minuti, attraverso un confronto tra la classificazione dello stato del traffico fatta sui valori osservati e su quelli previsti. Il traffico in ogni intervallo può essere congestionato o meno nella realtà e può essere ben previsto o meno dal modello. Il numero di falsi positivi e di falsi negativi risulta quindi essere l'indicatore più pertinente. Gli indicatori usati nella pratica sono:

- *Precision:*

$$Precision = \frac{V}{V + FP}$$

dove V indica gli intervalli che sono congestionati nella realtà e che sono previsti come tali, FP indica i falsi postivi, ovvero gli istanti che non sono congestionati nella realtà ma per i quali il modello ha previsto la congestione.

Recall:

$$Recall = \frac{V}{V + FN}$$

dove FN indica i falsi negativi, ovvero gli istanti che sono congestionati nella realtà ma per i quali il modello non ha previsto la congestione.

I risultati per i tre sensori analizzati in precedenza mostrano che l'algoritmo ha delle prestazioni migliori in *recall* che in *precision*: ci sono quindi meno falsi negativi (congestione non identificata) che falsi positivi (congestione identificata quando in realtà non sussiste). In ogni caso la precisione del modello è soddisfacente.

42+0250		
Congestione prevista	False	True
Congestione reale		
False	138 583	8 776
True	2 728	24 869

Precision	0,739
Recall	0,901

43+0900		
Congestione prevista	False	True
Congestione reale		
False	142 349	8 332
True	2 009	22 265

Precision	0,728
Recall	0,917

44+0970		
Congestione prevista	False	True
Congestione reale		
False	147 605	5 995
True	2 274	19 082

Precision	0,761
Recall	0,894

Tabella 10: Risultati del modello nell'individuazione della congestione – DiRIF

Occorre osservare che questi ultimi due indicatori derivano dalla definizione di congestione descritta nel paragrafo IV.D che corrisponde agli stati instabili del diagramma fondamentale (densità-flusso): tale definizione deriva tuttavia dall'ipotesi che sia identificato, valido e invariato il diagramma fondamentale, che in teoria dovrebbe essere applicabile solo a condizioni di traffico "time-indipendent" per un determinato tronco stradale. La prestazione del modello andrebbe quindi valutata con più attenzione sugli indicatori calcolati nei paragrafi precedenti. Infatti, dei valori non soddisfacenti di *precision* o di *recall* possono essere dovuti a una cattiva classificazione degli intervalli reali piuttosto che di quelli previsti.

C. Applicazione alla rete urbana di Lione

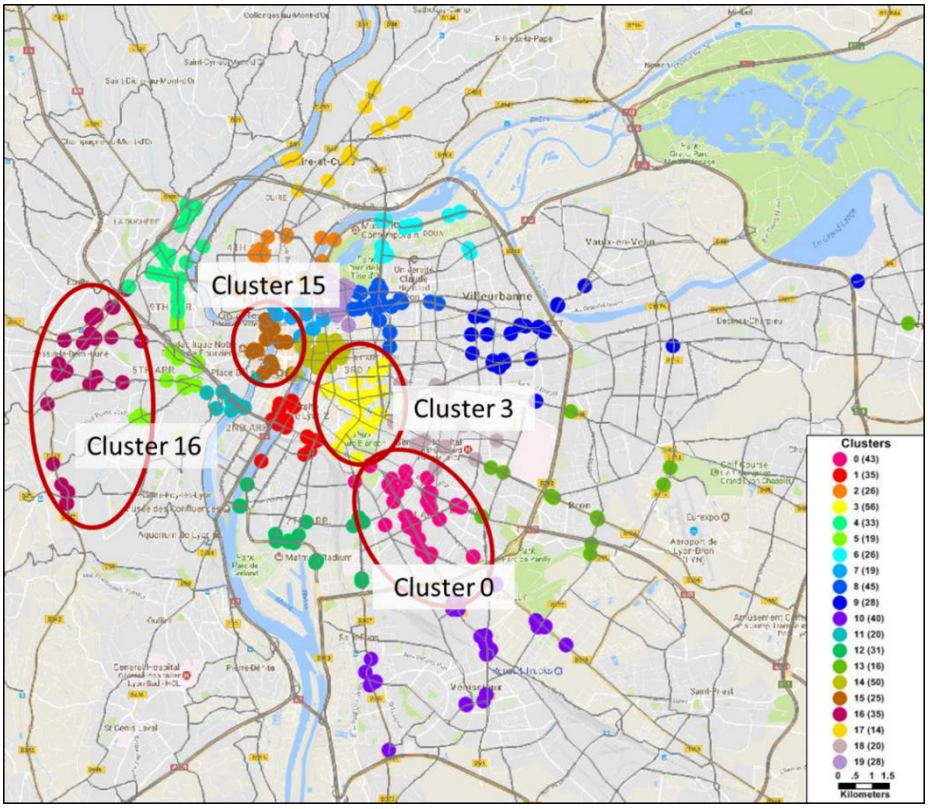


Figura 31: Clusters selezionati per l'analisi sulla rete urbana di Lione

L'analisi sulla rete urbana di Lione è stata fatta scegliendo come campione quattro diversi clusters (sui 20 totali). Questi cluster, presentati in Figura 17, hanno delle caratteristiche distinte tra loro: il cluster 0 si trova su uno degli assi di accesso alla città ed è caratterizzato da grandi viali lineari, il cluster 3 si trova in centro città, dove le strade formano un reticolo a maglie rettangolari, il cluster 15 si trova nel centro storico e turistico, vicino alla collina di Fourvière, infine il cluster 16 si trova ad ovest sulla D342, una strada dipartimentale che circonda la città. Per ogni cluster è stato scelto un segmento rappresentativo: su tale segmento si è tentato di prevedere la portata e il tasso di occupazione. Gli input del modello erano le rilevazioni ottenute in passato su tutti gli altri punti del cluster, oltre ovviamente alle misure sul segmento stesso.

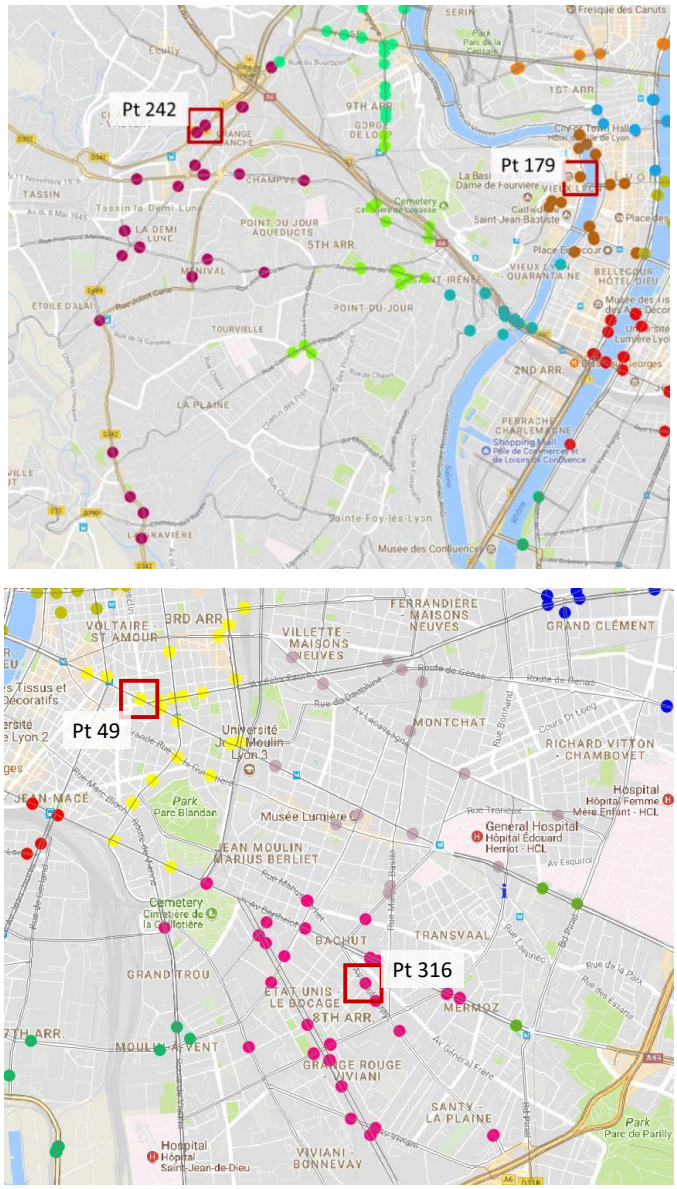


Figura 32: Localizzazione dei segmenti rappresentativi per i clusters studiati

I dati erano disponibili per tre anni (dal 2012 al 2014): per il test-set è stato scelto un periodo di 4 mesi. A causa di alcuni dati mancanti per il mese di dicembre 2014, il test-set va da agosto a novembre 2014. I risultati per 4 punti rappresentativi dei 4 cluster studiati sono illustrati in seguito.

1. Previsione della portata

Le misure dei sensori lionesi, nonostante siano state rilevate su intervalli di 6 minuti, sono state espresse in veicoli/ora (e non in veicoli per intervallo di 6 minuti come per i valori della DiRIF). Per poter comparare i risultati (in particolare gli errori) occorre dunque dividere per 10 le portate di Lione. Questa operazione non è stata effettuata a monte per ragioni tecniche e i risultati sono quindi presentati in veicoli/ora.

La qualità dei risultati per la previsione della portata totale è buona, con valori di r^2 sempre superiori a 80 e nella maggior parte dei casi prossimi a 90. Il comportamento delle serie previste è simile al caso extra-urbano.

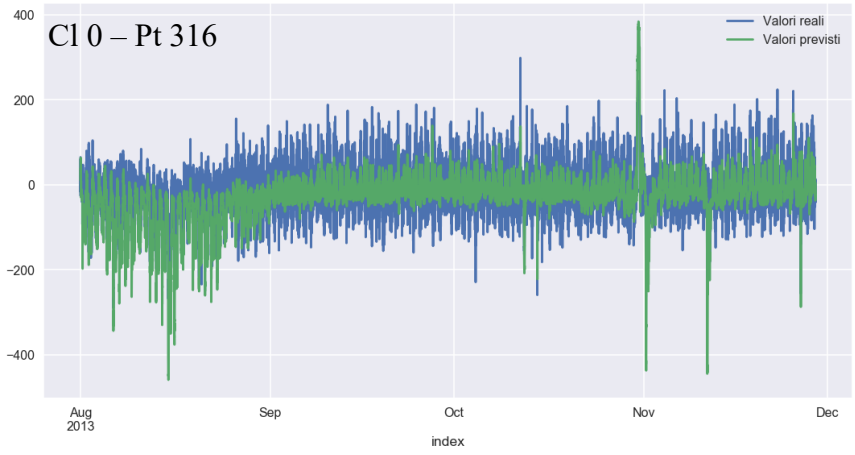
Cluster	Punto	R^2	RMSE	MAE
0	316	91,647	30,185	32,154
3	49	96,206	58,182	62,003
15	179	81,502	68,386	65,417
16	242	97,800	62,782	64,897

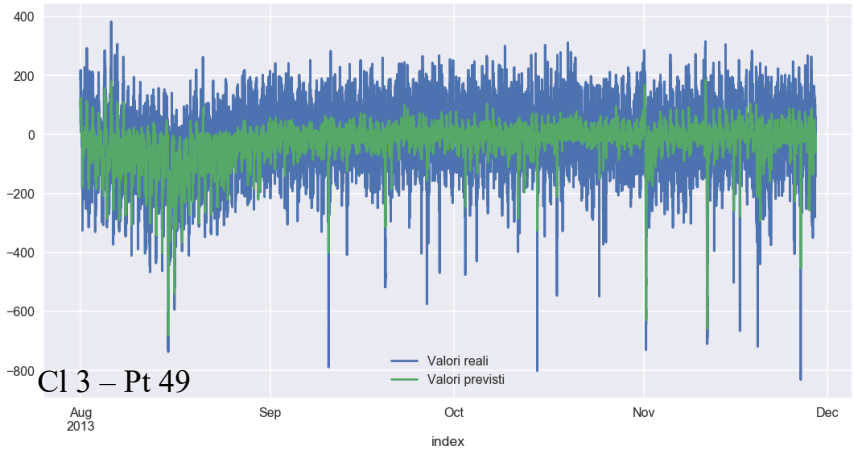
Tabella 11: Indicatori di performance – Lione – Portata totale

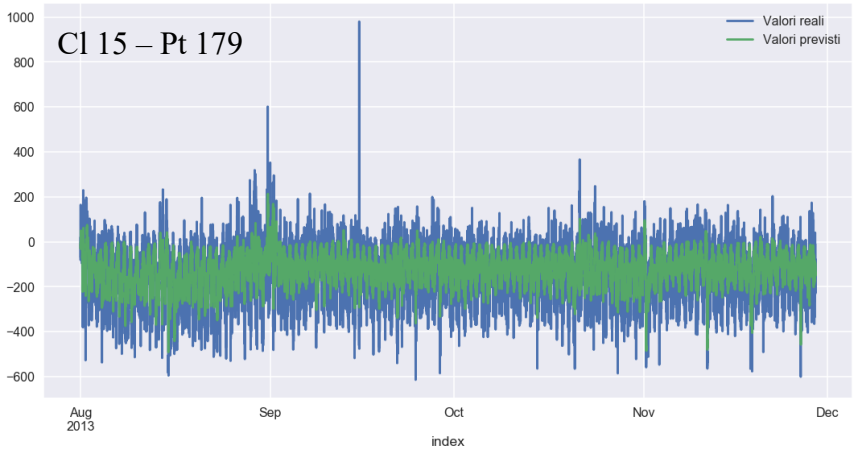
Se si passa ad analizzare la previsione del differenziale della portata, risulta evidente come i risultati siano meno buoni che nel caso extra-urbano. Il modello ha più difficoltà a prevedere la reale ampiezza delle oscillazioni e di conseguenza le sottostima. Questa sottostima è probabilmente dovuta alle caratteristiche della circolazione urbana e alla struttura della rete stradale che sono influenzate da fattori quali i semafori e le intersezioni o da comportamenti anomali degli automobilisti che, ad esempio, circolano a bassa velocità, non a causa della congestione ma per cercare parcheggio. Come nel caso della DiRIF, certe oscillazioni sono individuate con estrema precisione, in particolare nel mese di agosto e in corrispondenza delle vacanze.

Cluster	Punto	R^2	RMSE	MAE
0	316	55,773	42,141	30,019
3	49	36,900	80,281	59,350
15	179	45,226	86,100	63,973
16	242	65,921	87,182	62,284

Tabella 12: Indicatori di performance – Lione – Differenziale della portata







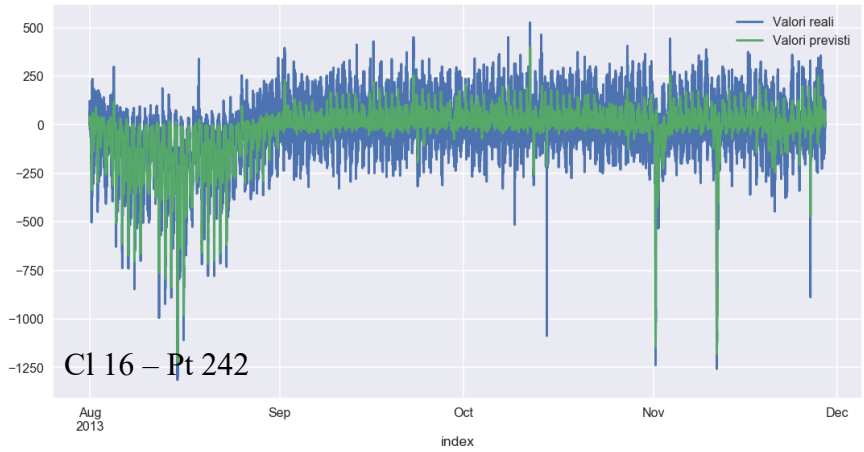
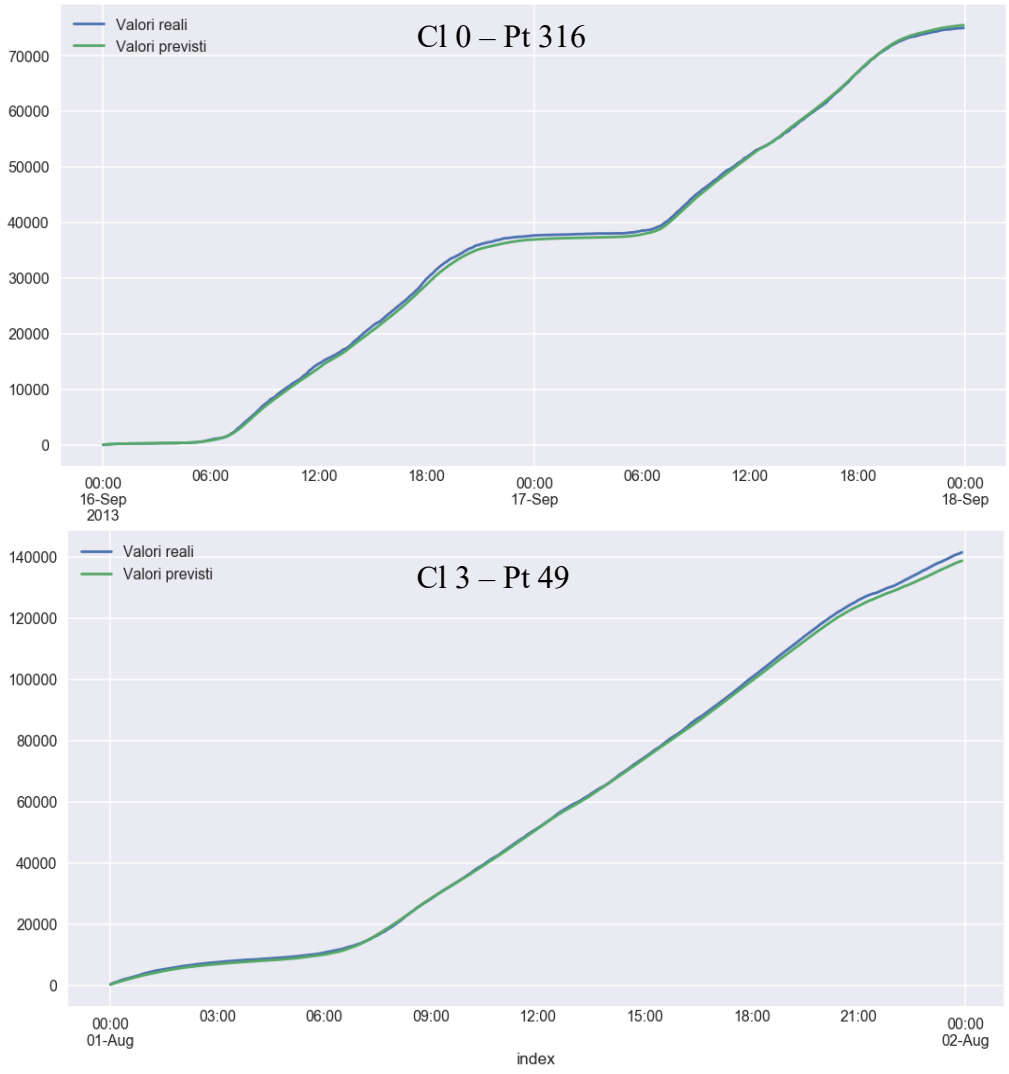


Figura 33: Rappresentazione grafica dei risultati per tutto il test-set – Lione – Differenziale della portata (veh/h in 6 min)



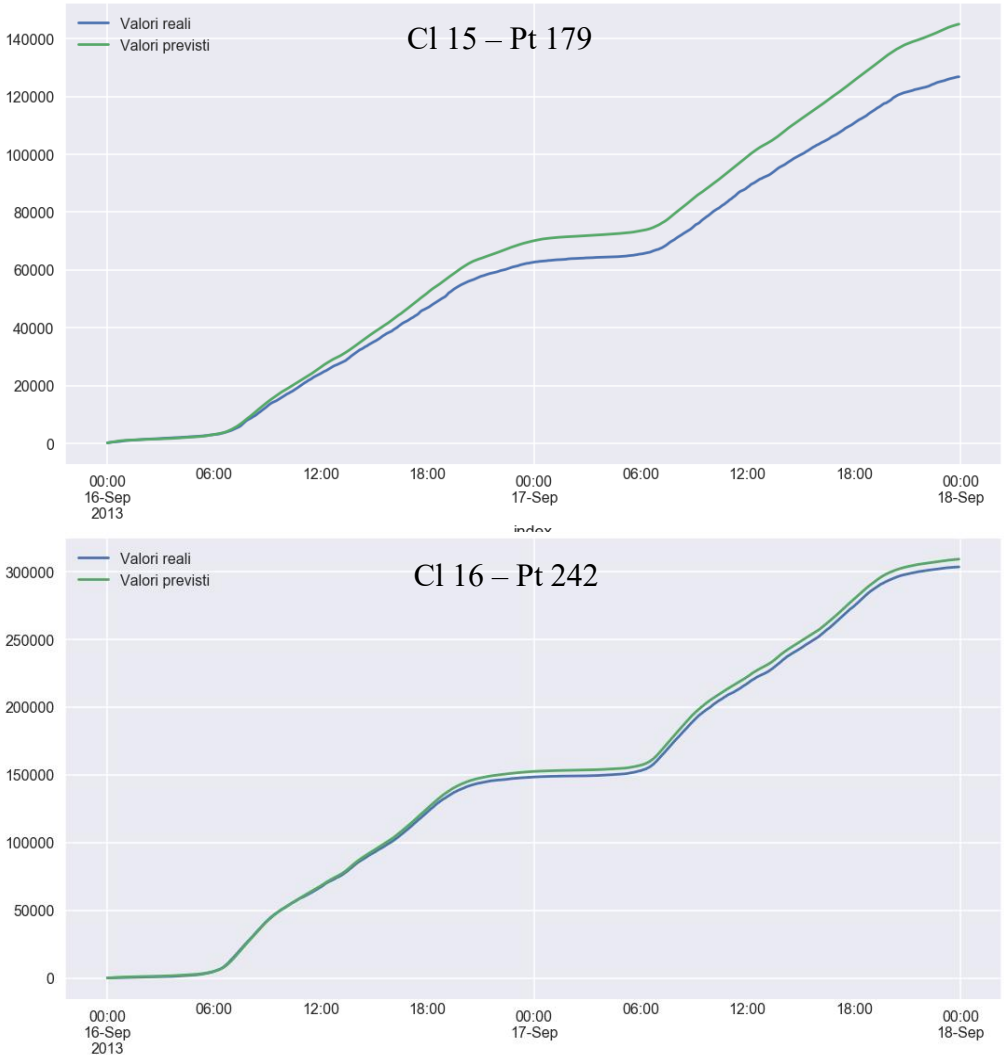
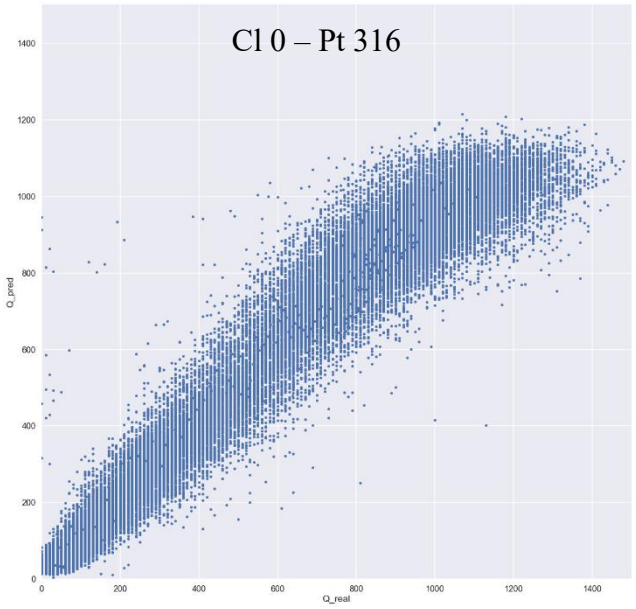


Figura 34: Rappresentazione dei risultati in forma cumulata – Lione – Portata totale (veh/h in 6 min)



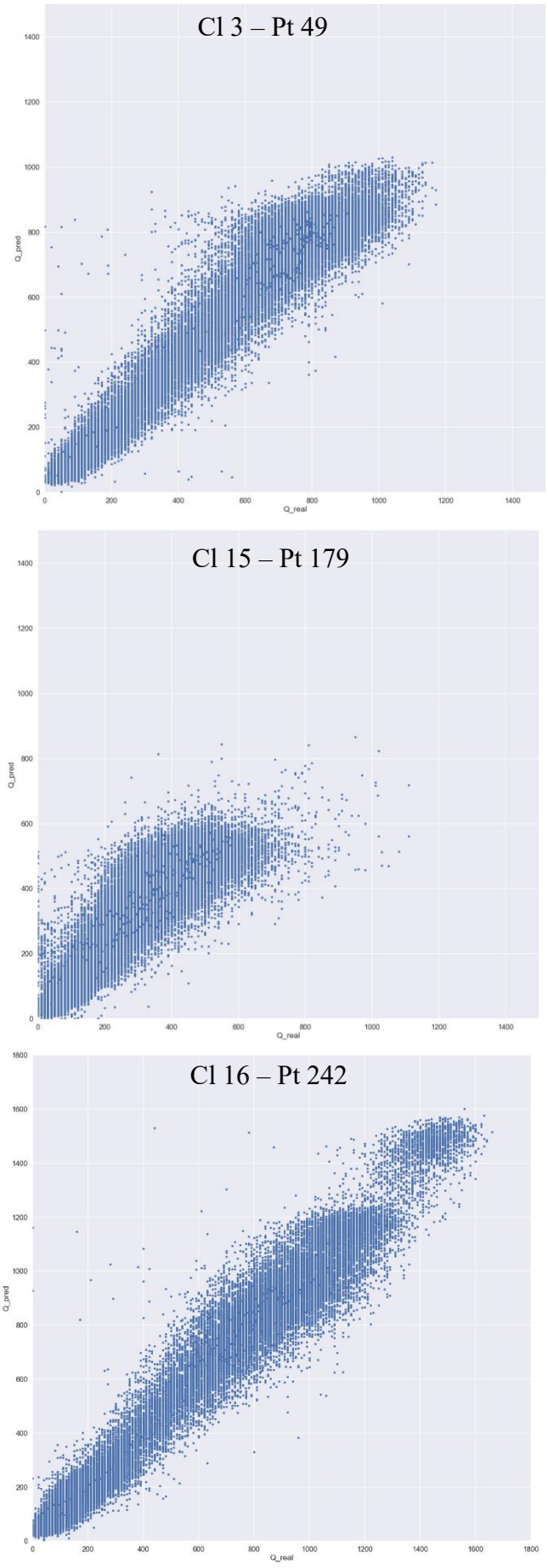


Figura 35: Correlazione tra valori osservati e valori previsti – Lione – Portata totale

La correlazione tra valori osservati e valori previsti conferma la tendenza del modello a sottostimare le portate quando queste sono molto grandi, come già osservato in ambito extra-urbano. I risultati evidenziano anche, in generale, una minore precisione nel caso urbano rispetto a quanto osservato nel caso extra-urbano. I risultati peggiori si ottengono per i cluster 3 e 15, che si trovano in pieno centro città, mentre i migliori si ottengono per il cluster 16 che si trova su una strada con un regime più simile a quello extra-urbano. I grafici delle portate cumulate e della correlazione per il cluster 15 mostrano come, durante il periodo di test, il traffico su questo segmento abbia avuto un comportamento anomalo. In seguito ad un'attenta analisi della serie storica, ci si accorge che i valori di portata per il sensore subiscono un brusco abbassamento dopo l'estate 2013, ovvero in corrispondenza del test-set. I risultati su questo sensore assumono allora interesse particolare, in quanto diventa possibile valutare la robustezza del modello di fronte a fenomeni imprevisti

Analizzando l'RMSE e l'MAE, dopo averli divisi per 10 per il motivo spiegato in precedenza, ci si accorge che sono dell'ordine dei 10 veicoli, valore che corrisponde all'incirca a un errore del 5 %.

2. Previsione del tasso di occupazione

Come per la portata, anche per il tasso di occupazione i risultati sono meno buoni che nel caso extra-urbano: l'r^2 è più basso e gli errori sono dello stesso ordine di grandezza del caso della DiRIF, nonostante il flusso totale di veicoli sia minore.

Cluster	Punto	R²	RMSE	MAE
0	316	58,334	6,052	3,057
3	49	76,381	1,687	0,753
15	179	21,148	3,429	1,320
16	242	95,638	0,895	0,500

Tabella 13: Indicatori di performance – Lione – Tasso di occupazione

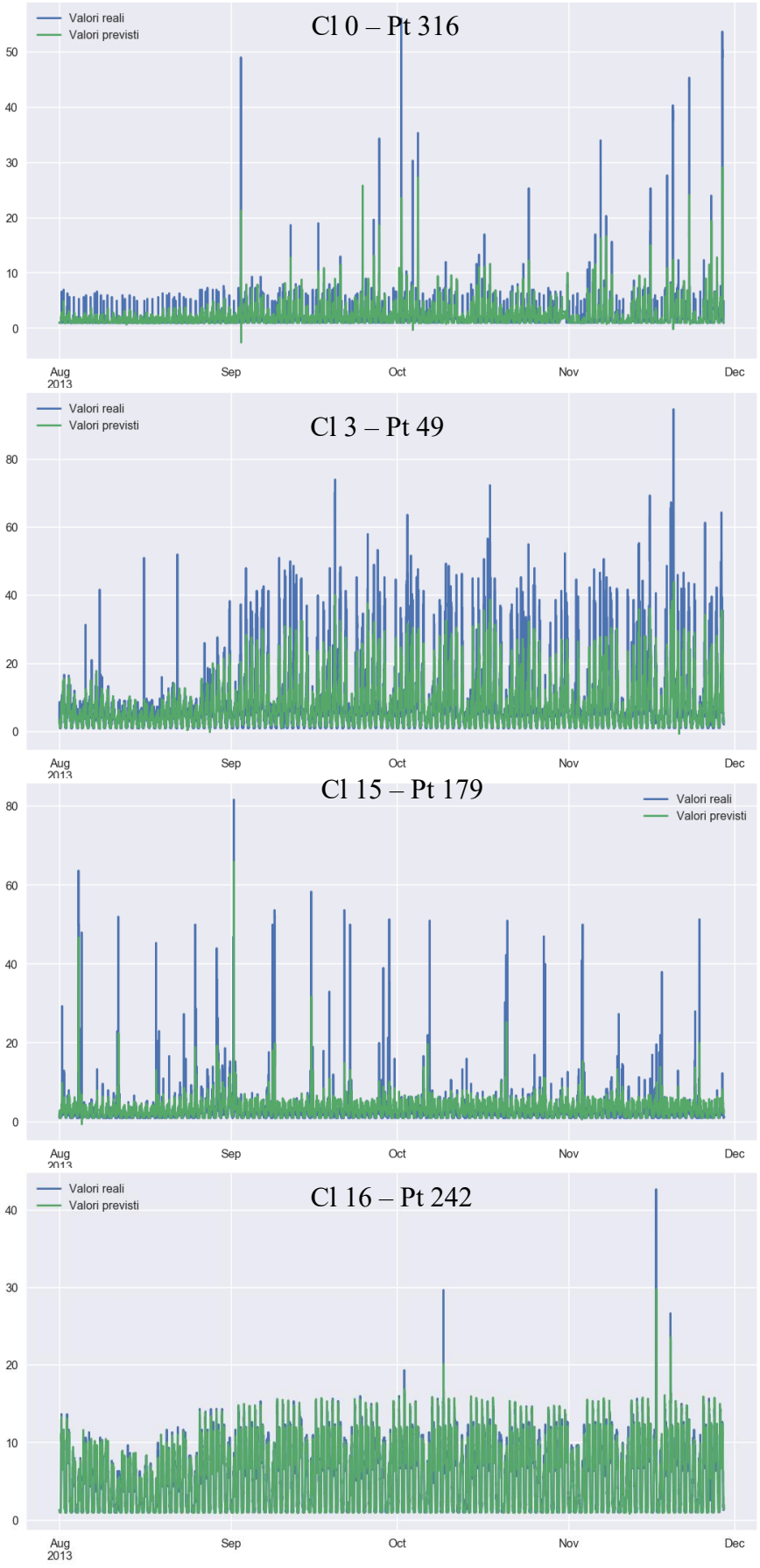


Figura 36: Rappresentazione grafica dei risultati per tutto il test-set – Lione – Tasso di occupazione

Anche nel caso del tasso di occupazione, è evidente come le performance del modello in ambito urbano siano molto diverse a seconda del cluster considerato: in particolare i risultati sono migliori per i clusters 0 e 16, che sono più lineari e periferici.

3. Individuazione della congestione

I risultati confermano anche in questo caso che il modello è migliore in recall che in precision. Inoltre, la previsione è ancora una volta migliore nel caso extra-urbano che in quello urbano e i risultati sono migliori, come nel caso precedente, per i cluster 0 e 16.

Pt 316		
Congestione prevista	False	True
Congestione reale		
False	233 293	8 643
True	2 880	17 980

Precision	0,675
Recall	0,862

Pt 49		
Congestione prevista	False	True
Congestione reale		
False	217 301	18 835
True	5 897	20 763

Precision	0,524
Recall	0,779

Pt 179		
Congestione prevista	False	True
Congestione reale		
False	240 251	9 201
True	2 149	11 195

Precision	0,549
Recall	0,839

Pt 242		
Congestione prevista	False	True
Congestione reale		
False	233 622	7 989
True	2 648	18 537

Precision	0,699
Recall	0,875

Tabella 14: Risultati del modello nell'individuazione della congestione – Lione

D. Valutazione dei risultati rispetto alla previsione con ipotesi semplificate

Per meglio comprendere la qualità delle previsioni del modello si è scelto di confrontare i risultati di portata e tasso di occupazione ottenuti dall'applicazione dell'algoritmo GBR con quelli che si otterrebbero tramite approcci semplificati, ovvero senza l'uso di algoritmi di auto-apprendimento.

Questi test permettono di apprezzare la potenzialità dell'approccio *data-driven* e di valutare il reale contributo alla previsione apportato dal machine learning. Il confronto è realizzato attraverso la comparazione degli indicatori di performance calcolati nei paragrafi V.B e V.C con quelli ottenuti dagli approcci semplificati per tutti i punti precedentemente analizzati sia nel caso urbano che nel caso extra-urbano.

1. Confronto con il valor medio delle osservazioni storiche

Il primo approccio semplificato consiste nell'assumere che, per un determinato istante della giornata, le condizioni di traffico siano uguali al valor medio della serie storica. Ciò equivale ad una condizione estrema in cui la tecnica di previsione non prende in considerazione i dati in tempo reale: la portata e il tasso di occupazione sono determinati esclusivamente a partire dalle osservazioni passate.

In questo confronto i periodi coperti dal train-set e dal test-set sono identici a quelli già introdotti nei paragrafi precedenti: per il caso urbano il train-set va da gennaio 2012 a luglio 2014 e il test-set da agosto a novembre 2014, per il caso extra-urbano il train-set va da gennaio 2016 ad ottobre 2017 e il test-set da novembre a dicembre 2017.

Per ognuno dei 1680 intervalli di sei minuti della settimana i rispettivi valori di portata e tasso di occupazione sono stati calcolati come media di tutte le misure storiche del train-set. Sono stati quindi ricavati gli indicatori di performance relativi a questo approccio e successivamente confrontati con quelli ottenuti dall'applicazione dell'algoritmo GBR già presentati nei paragrafi precedenti.

a. Rete extra-urbana della DiRIF

La qualità dei risultati ottenuti con questo primo approccio semplificato in ambito extra-urbano è di gran lunga inferiore a quelli ottenuti dall'applicazione del modello e già presentati nel paragrafo V.B.

Nel caso della portata (Tabella 15) il valore di r^2 passa da valori superiori a 95 a valori compresi tra 55 e 50, mentre sia l'RMSE che l'MAE errori sono tre volte più importanti.

Risultati modello				Metodo semplificato			
Progressiva	R^2	RMSE	MAE	Progressiva	R^2	RMSE	MAE
42+0250	95,918	21,118	14,582	42+0250	54,819	71,279	46,279
43+0900	96,527	19,891	13,704	43+0900	50,355	74,671	50,062
44+0970	96,811	18,823	12,836	44+0970	52,138	72,936	47,421

Tabella 15: Confronto tra i risultati del modello e i risultati del primo metodo semplificato - DiRIF – Portata totale

Nel caso del tasso di occupazione (Tabella 16) i risultati sono ancora meno buoni, a causa del carattere estremamente variabile di questa grandezza che mal si presta ad essere descritta esclusivamente con il valor medio. I valori di r^2 sono quasi tre volte inferiori a quelli ottenuti con l'applicazione del modello, mentre gli errori sono due volte più grandi. È importante osservare che il tasso di occupazione assume, per la maggior parte del tempo, valori inferiori al 20%: un errore dell'ordine del 6-8% diventa quindi molto rilevante per la previsione.

Risultati modello				Metodo semplificato			
Progressiva	R^2	RMSE	MAE	Progressiva	R^2	RMSE	MAE
42+0250	84,931	3,688	1,690	42+0250	30,341	6,467	4,144
43+0900	87,850	2,432	1,204	43+0900	36,893	5,710	3,652
44+0970	79,534	3,505	1,914	44+0970	29,911	8,105	4,829

Tabella 16: Confronto tra i risultati del modello e i risultati del primo metodo semplificato - DiRIF – Tasso di occupazione

a. Rete urbana di Lione

In ambito urbano l'approccio semplificato fornisce risultati diversi a seconda del cluster e della variabile considerate.

Per la previsione della portata (Tabella 17), la differenza di prestazione rispetto al caso extra-urbano è meno marcata sia in termini di r^2 che in termini di errore. Il valore di r^2 negativo per il punto del cluster 15, dovuto ad un improvviso abbassamento dei valori di portata sul tratto stradale all'inizio del test-set, mostra come una previsione basata esclusivamente sui valori medi non garantisce alcuna robustezza in caso di eventi eccezionali. Al contrario il modello in tempo reale, benché le sue prestazioni risentano del cambiamento di comportamento del traffico, permette di ottenere delle previsioni comunque affidabili.

Risultati modello					Metodo semplificato				
Cluster	Punto	R^2	RMSE	MAE	Cluster	Punto	R^2	RMSE	MAE
0	316	91,647	30,185	32,154	0	316	64,310	64,934	44,154
3	49	96,206	58,182	62,003	3	49	88,607	102,846	71,819
15	179	81,502	68,386	65,417	15	179	-17,338	181,699	147,961
16	242	97,800	62,782	64,897	16	242	87,607	150,761	92,246

Tabella 17: Confronto tra i risultati del modello e i risultati del primo metodo semplificato - Lione – Portata totale

Risultati modello					Metodo semplificato				
Cluster	Punto	R^2	RMSE	MAE	Cluster	Punto	R^2	RMSE	MAE
0	316	58,334	6,052	3,057	0	316	8,115	2,289	1,039
3	49	76,381	1,687	0,753	3	49	38,590	7,483	3,842
15	179	21,148	3,429	1,320	15	179	14,914	4,840	1,484
16	242	95,638	0,895	0,500	16	242	83,328	1,782	1,111

Tabella 18: Confronto tra i risultati del modello e i risultati del primo metodo semplificato - Lione – Tasso di occupazione

2. Confronto con i valori correnti assunti costanti per 30 minuti

Il secondo approccio semplificato consiste nell'assumere che i valori di portata e tasso di occupazione osservati all'istante t siano validi per tutto l'intervallo da t a $t + n$. Ciò equivale ad una condizione estrema in cui la tecnica di previsione non prende in considerazione la serie storica ma si basa esclusivamente sui dati osservati in tempo reale.

Si è quindi ipotizzato di disporre solamente di una misura ogni mezz'ora e di assumerla costante per i successivi 30 minuti, successivamente una nuova misura è disponibile e il processo viene ripetuto iterativamente per tutto il test-set. Sono stati quindi calcolati gli indicatori di performance relativi a questo approccio e successivamente confrontati con quelli ottenuti dall'applicazione dell'algoritmo GBR e già presentati nei paragrafi precedenti. Anche in questo caso il train-set e il test-set coincidono con quelli già utilizzati in precedenza.

a. Rete extra-urbana della DiRIF

La qualità dei risultati di questo secondo approccio semplificato è, in ambito extra-urbano, migliore rispetto al caso analizzato in precedenza, ma sempre inferiore a quella offerta dal modello.

Per la portata totale (Tabella 19), l'r^2 ottenuto assumendo i valori correnti come validi per 30 minuti è dell'ordine di 90, contro 95 delle previsioni del modello, mentre gli errori commessi sono più grandi del 50% circa. Analizzando la Tabella 20, relativa al differenziale della portata,

si osserva come il peggioramento della qualità dei risultati sia più importante nel caso della previsione con ipotesi semplificate.

Risultati modello				Metodo semplificato			
Progressiva	R^2	RMSE	MAE	Progressiva	R^2	RMSE	MAE
42+0250	95,918	21,118	14,582	42+0250	91,684	30,583	19,701
43+0900	96,527	19,891	13,704	43+0900	89,370	34,555	22,085
44+0970	96,811	18,823	12,836	44+0970	89,526	34,120	21,400

Tabella 19: Confronto tra i risultati del modello e i risultati del secondo metodo semplificato - DiRIF – Portata totale

Risultati modello				Metodo semplificato			
Progressiva	R^2	RMSE	MAE	Progressiva	R^2	RMSE	MAE
42+0250	88,690	20,851	14,218	42+0250	81,594	30,427	19,581
43+0900	87,506	19,891	13,704	43+0900	78,588	34,321	21,999
44+0970	90,667	21,776	15,285	44+0970	78,012	34,014	21,391

Tabella 20: Confronto tra i risultati del modello e i risultati del secondo metodo semplificato - DiRIF – Differenziale della portata[5]

Anche nel caso del tasso di occupazione la qualità della previsione con questo metodo semplificato è migliore che nel caso del valor medio ma inferiore a quella del modello.

Risultati modello				Metodo semplificato			
Progressiva	R^2	RMSE	MAE	Progressiva	R^2	RMSE	MAE
42+0250	84,931	3,688	1,690	42+0250	66,450	4,488	2,118
43+0900	87,850	2,432	1,204	43+0900	75,701	3,544	1,756
44+0970	79,534	3,505	1,914	44+0970	67,508	5,519	2,358

Tabella 21: Confronto tra i risultati del modello e i risultati del secondo metodo semplificato - DiRIF – Tasso di occupazione

b. Rete urbana di Lione

Le osservazioni fatte per il caso extra-urbano restano valide anche in ambito urbano. Il secondo approccio semplificato fornisce risultati migliori rispetto al primo, ma meno buoni rispetto al modello, sia per la portata che per il tasso di occupazione.

[5] Nel caso del primo approccio semplificato (valor medio delle osservazioni storiche) non sono stati calcolati gli indicatori di performance relativi al differenziale della portata. Questo perché, secondo la scomposizione delle serie storiche fatta dal modello e descritta nel paragrafo IV.B.a, i valori medi delle osservazioni storiche coincidono esattamente con la stagionalità.

Risultati modello					Metodo semplificato				
Cluster	Punto	R^2	RMSE	MAE	Cluster	Punto	R^2	RMSE	MAE
0	316	91,647	30,185	32,154	0	316	72,842	56,645	39,582
3	49	96,206	58,182	62,003	3	49	83,352	124,329	92,445
15	179	81,502	68,386	65,417	15	179	58,932	107,501	79,271
16	242	97,800	62,782	64,897	16	242	85,095	165,337	114,260

Tabella 22: Confronto tra i risultati del modello e i risultati del secondo metodo semplificato - Lione – Portata totale

Risultati modello					Metodo semplificato				
Cluster	Punto	R^2	RMSE	MAE	Cluster	Punto	R^2	RMSE	MAE
0	316	55,773	42,141	30,019	0	316	20,093	61,645	38,157
3	49	36,900	80,281	59,350	3	49	17,976	91,513	67,843
15	179	45,226	86,100	63,973	15	179	42,686	88,073	65,198
16	242	65,921	87,182	62,284	16	242	44,928	110,834	76,886

Tabella 23: Confronto tra i risultati del modello e i risultati del secondo metodo semplificato - Lione – Differenziale della portata

Risultati modello					Metodo semplificato				
Cluster	Punto	R^2	RMSE	MAE	Cluster	Punto	R^2	RMSE	MAE
0	316	58,334	6,052	3,057	0	316	19,567	2,141	0,649
3	49	76,381	1,687	0,753	3	49	11,311	8,993	4,160
15	179	21,148	3,429	1,320	15	179	16,599	3,954	1,454
16	242	95,638	0,895	0,500	16	242	80,288	1,937	1,122

Tabella 24: Confronto tra i risultati del modello e i risultati del secondo metodo semplificato - Lione – Tasso di occupazione

E. Possibili sviluppi dello studio

I risultati della sperimentazione sono stati incoraggianti e hanno permesso di individuare delle azioni di miglioramento delle procedure di stima. Alcune di queste sono già state indicate nei precedenti paragrafi, come ad esempio il trattamento dei dati mancanti e l'approfondimento della definizione delle variabili.

Il trattamento dei dati mancanti potrà sicuramente apportare dei miglioramenti nella performance della previsione. L'interpolazione lineare dei dati mancanti, adottata come soluzione nel presente studio, introduce un'eccessiva semplificazione e un errore sistematico al momento dell'adattamento del modello e del calcolo delle performances. Infatti, l'allenamento dell'algoritmo su una serie contenente delle interpolazioni causa una perdita di informazioni al

momento dell'apprendimento che, come detto, ha come conseguenza una perdita di performance del modello. Inoltre, nel calcolo degli indicatori r^2, RMSE e MAE questa approssimazione causa una sottostima delle reali performances. Se questi valori mancanti della serie fossero ricostruiti con maggiore precisione, gli errori sarebbero probabilmente più piccoli e quindi le performances risulterebbero migliori.

Come anticipato nel paragrafo IV.B, sarà sicuramente necessario apportare delle migliorie nella definizione delle variabili descrittive del modello. La variabile booleana "Heure de pointe", che indica se l'istante considerato fa parte del periodo della giornata in cui il traffico è più importante, è stata abbandonata perché poco significativa per il modello. Tenendo conto che l'ora di punta è una caratteristica intrinseca del traffico, e di conseguenza molto importante, è assai probabile che la debole importanza di tale variabile sia dovuta ad una sua definizione troppo approssimativa. In questa sperimentazione la variabile "Heure de pointe" è stata definita a partire dall'istante della giornata in cui il traffico era massimo: l'ora di punta comprendeva la mezz'ora precedente e la mezz'ora seguente a quell'istante. Nella realtà, soprattutto in ambito urbano, è probabile che una singola ora di punta non sia sufficiente a descrivere il fenomeno. Si può, per esempio, immaginare la creazione di due variabili ora di punta, una al mattino e una alla sera, e tramite una definizione che prenda in conto anche il tasso di occupazione. In ogni caso anche una variabile così definita necessiterebbe di essere introdotta e testata per verificare che il suo contributo sia veramente significativo per la prestazione del modello.

Le variabili meteorologiche "Pluie" e "Neige" sono state anch'esse trascurate a causa della loro debole importanza. I dati di meteo France sono disponibili per intervalli di 3 ore e si tratta di un unico valore per tutta la città, quindi probabilmente non abbastanza "fini" per una previsione puntuale con intervalli di 6 minuti. Ciononostante, soprattutto per la variabile "Pluie", che ha una varianza maggiore, sarà probabilmente necessario rivalutare la sua definizione. È probabile che una variabile booleana definita a partire da un valore limite arbitrario di pioggia non sia sufficiente per descrivere un fenomeno complesso come la congestione stradale. Sarà probabilmente necessario creare più classi di precipitazione secondo l'intensità della pioggia. Questo potrebbe rendere la variabile più significativa per il modello.

VI. Conclusioni

Questa tesi ha avuto l'obiettivo di sviluppare e testare uno strumento per la previsione in tempo reale della congestione stradale, applicabile ad un orizzonte temporale di 30 minuti. Tale strumento è stato testato sia su tronchi stradali urbani che extra-urbani.

Nella prima parte dello studio si sono analizzati e confrontati i potenziali dati di input: misure di portata e tasso di occupazione rilevati tramite sensori a postazione fissa (spire induttive) e misure di velocità media rilevate tramite sensori su veicoli (Floating Car Data - FCD). In seguito, concentrandosi sulle misure delle spire induttive, i dati di input sono stati analizzati ed elaborati con un duplice obiettivo: da un lato la definizione di variabili temporali, spaziali e metereologiche per alimentare il modello, dall'altro la disaggregazione delle serie osservate in una componente costante, che non necessita di essere predetta, ed una componente variabile, oggetto della previsione. Attraverso l'applicazione di un algoritmo di machine learning alle variabili precedentemente definite si sono ottenute previsioni dei valori di flusso di traffico e tasso di occupazione del segmento in esame ad un orizzonte temporale di 30 minuti; tali valori sono stati confrontati con le misure realmente osservate. Infine si è tentato, a partire dai valori predetti, di determinare se tali condizioni di traffico comportino o meno l'insorgere di fenomeni locali di congestione stradale.

I risultati ottenuti da questa esperienza si possono considerare soddisfacenti sotto molti punti di vista.

Il confronto con le previsioni tramite approcci semplificati ha messo in evidenza la buona qualità dei risultati ottenuti con il modello sperimentale e dimostra che la previsione della congestione può essere affrontata anche attraverso l'approccio *data-driven*. L'interesse di questo approccio risiede nella possibilità di applicare all'ingegneria del traffico gli algoritmi di *machine learning*: tali algoritmi permettono di sfruttare una mole sempre crescente di dati di mobilità di cui disponiamo oggi, senza la necessità di costruire modelli matematici che tentano di riprodurre gli aspetti più complessi di comportamento del sistema.

Le caratteristiche migliori dei *big data* si sono rivelate essere il loro volume e la loro disponibilità. L'alimentazione di un modello predittivo del traffico è possibile solo se si ha a disposizione una grande quantità di dati di input che descrivano i principali parametri del

traffico. Inoltre, affinché il modello possa funzionare in tempo reale e in continuo, i dati devono essere disponibili senza interruzioni su tutto l'arco temporale di osservazione. Di contro, i big data presentano ancora problemi in termini di affidabilità. Questo è vero soprattutto per le velocità FCD che, pur rappresentando la risorsa che presumibilmente sarà più utilizzata in futuro, richiedono, prima di essere adoperate, uno studio approfondito per ben comprenderne le caratteristiche e di conseguenza le potenzialità.

A conclusione del lavoro svolto in questo progetto, è possibile affermare che la previsione delle caratteristiche principali di un flusso di traffico ad un orizzonte di 30 minuti può essere effettuata con una buona precisione per entrambi gli ambiti studiati. Nei test sulla rete extra-urbana dell'Ile-de-France la portata è stata prevista con un r^2 del 95% e un errore medio inferiore al 10% e il tasso di occupazione con un r^2 compreso tra l'80 e il 90%. Nei test sulla rete urbana di Lione si è osservata una maggiore variabilità nei risultati: per la portata l'r^2 è compreso tra l'80 e il 97%, per il tasso di occupazione l'r^2 oscilla dal 20 al 95%. In generale la previsione della portata e del tasso di occupazione si sono rivelate migliori in ambito extra-urbano che in quello urbano. Quest'ultimo risultato evidenzia come il modello sia più efficiente nella previsione del traffico quando questo è in condizioni più prossime al deflusso ininterrotto e quando la struttura della rete stradale è semplice con pochi nodi e archi.

Come sottolineato nel paragrafo V.B.3, i risultati dell'individuazione della congestione dipendono fortemente dalla definizione stessa di congestione che si introduce nel modello. Identici valori di portata e densità su due infrastrutture differenti possono dare origine a condizioni di traffico diverse, che possono essere descritte da specifici diagrammi fondamentali, costruiti per rappresentare gli stati del traffico di un particolare tronco stradale. La definizione della congestione non è quindi univoca per i vari tronchi di un'arteria. Un modello in grado di funzionare con molteplici definizioni di congestione potrà quindi applicarsi a contesti differenti grazie alla sua maggiore versatilità.

Infine, le azioni di miglioramento descritte in precedenza, se adeguatamente testate, consentiranno in futuro di affinare il modello e di aumentarne l'affidabilità e la precisione della risposta.

Le buone performance ottenute fin dalla prima fase di sperimentazione del modello inducono interesse per la sua ottimizzazione, attraverso le azioni di miglioramento ipotizzate, e per la sua necessaria generalizzazione: entrambi gli obiettivi potranno essere raggiunti con un'estesa

sperimentazione di tipo applicativo. Gli sforzi in questa direzione sono giustificati perché una migliore previsione in tempo reale della congestione stradale può apportare un importante contributo alla gestione del traffico, che ha un impatto importante sulla vivibilità delle grandi metropoli, sia in termini sociali che in termini economici.

VII. Bibliografia

[1] INRIX, «INRIX Global Traffic Scorecard,» February 2018.

[2] P.-A. Laharotte, *Contributions à la prévision court-terme, multi-échelle et multi-variée, par apprentissage statistique du trafic routier, thèse de doctorat, spécialité: genie civil,* Lyon: Université de Lyon, 2016.

[3] D. Robinson, «The incredible growth of Python,» 6 September 2017. [Online]. Available: https://stackoverflow.blog/2017/09/06/incredible-growth-python/. [Consultato il giorno Mai 2018].

[4] «Better traffic predictions,» Datacity, [Online]. Available: https://www.datacity.numa.co/single-post/Better-traffic-predictions. [Consultato il giorno Juillet 2018].

[5] A. Caragliu, C. Del Bo e P. Nijkamp, *Smart cities in Europe,* VU University Amsterdam: Faculty of Economics, Business Administration and Econometrics, 2009.

[6] «Ville intelligente,» Wikipédia, [En ligne]. Available: https://fr.wikipedia.org/wiki/Ville_intelligente. [Accès le Juin 2018].

[7] S. Cohen e M. Danech-Pajouh, *Initiation à l'ingenierie du trafic,* 2000.

[8] B. D. Greenshields, «A study of traffic capacity,» *Highway Research Board Processing, Volume 14,* December 1934.

[9] H. Greenberg, «An Analysis of Traffic Flow,» *Operations Research ,Volume 7 Issue 1,* pp. 79-85, February 1959.

[10] S. Cohen, Ingénierie du trafic routier, Paris: Presses de l'école nationale des Ponts et chaussées, 1993.

[11] B. D. Greenshields, «The photographic method of studying traffic behavior,» *Highway Research Board Proceedings,* vol. 13, 1934.

[12] S. Maerivoet e B. De Moor, «Transportation Planning and Traffic Flow Models,» Leuven (BE), 2 february 2008.

[13] H. J. Payne, «FREFLO: A Macroscopic Simulation Model of Freeway Traffic,» *Transportation Research Record,* n. 722, pp. 68-77, 1979.

[14] R. Jiang, Q. Wu e Z. Zhu, «Full velocity difference model for a carfollowing theory,» *Physical Review E,* vol. 64, n. 1, 2001.

[15] R. O. Duda, P. E. Hart e D. G. Stork, Pattern classification, John Wiley & Sons, 2012.

[16] M. Bugdol, Z. Miodonska, M. Krecichwost e P. Kasperek, «Vehicle detection system using magnetic sensors,» *Transport Problems,* vol. 9, n. 1, pp. 49-60, 2014.

[17] A. Mocholi-Salcedo, J. H. Arroyo-Nunez, V. Milian-Sanchez, A. Arroyo-Nunez e G. J. Verdú-Martín, «Traffic Control Magnetic Loops Electric Characteristics Variation Due to the Passage of Vehicles Over Them,» *IEEE Transactions on Intelligent Transportation Systems,* vol. 18, n. 6, pp. 1540-1548, 2017.

[18] «Traffic Detector Handbook - Third edition Volume I,» US Department of Transportation - Federal Highway Administration, October 2006.

[19] Lyon Métropole, «DATA Grand Lyon,» [Online]. Available: https://data.grandlyon.com/search/?Q=criter. [Consultato il giorno Avril 2018].

[20] R. Remesy, «Floating Car Data : quel bilan pour la gestion du trafic?,» *TEC Magazine,* n. 237, pp. 38-39, Avril 2018.

[21] A. Anselmi, P. Chiodini e F. Verrecchia, «Données manquantes et prévisions: Méthodes à imputation variable,» Paris, 2009.

[22] G. Melard, «Modelisation de series temporelles a haute frequence : aspects logiciels,» ECARES, Universite libre de Bruxelles, Bruxelles.

[23] J. Perktold, S. Seabold e J. Taylor, «Statsmodel : Statistics in Python,» Statsmodel developers, [Online]. Available: http://www.statsmodels.org/dev/index.html. [Consultato il giorno Juin 2018].

[24] L. Breiman, J. Friedman, C. Stone e R. Olsen, Classification and Regression Trees, Taylor & Francis, 1984.

[25] L. Hawarah, Une approche probabiliste pour le classement d'objets, Université Joseph-Fourier - Grenoble I: Informatique [cs], 2008.

[26] «Données publiques,» Météo France, [Online]. Available: https://donneespubliques.meteofrance.fr. [Consultato il giorno Juin 2018].

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot e É. Duchesnay, «Scikit-learn: Machine Learning in Python,» *Journal of Machine Learning Research,* vol. 12, pp. 2825--2830, 2011.