UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

Dipartimento di INFORMATICA

Dottorato di ricerca in Informatica
XXVII ciclo

Progetto di ricerca

Dottorando: Dott.ssa Sonja Pravilovic

Tutor: Dott.ssa Annalisa Appice

Coordinatore

Prof. Donato Malerba

Firma del dottorando: _____

Firma del tutor: _____

INDICE

1) Titolo della ricerca:

Sintesi di tecniche di predizione per serie temporali geo-riferite.

2) Area nella quale si inquadra la ricerca:

Analisi di serie temporali, Data mining spaziale, temporale e spazio-temporale, Stream data mining.

3) Obiettivi della ricerca

L'obiettivo della ricerca è quello di investigare nuove teorie, tecniche e algoritmi per la estrazione di conoscenza da serie temporali geo-riferite e utilizzare tale conoscenza per costruire modelli di previsione "accurati" che si adattino ai naturali cambiamenti, in spazio e tempo, attesi nella evoluzione di campi geo-fisici.

In letteratura, sono già riportate alcune tecniche di data mining che apprendono e sperimentano modelli di previsione appresi da dati spazio-temporali. L'obiettivo di questa ricerca è estendere tali soluzioni o sintetizzarne di nuove che opportunamente modellino le differenti forme di correlazione (nel tempo e/o nello spazio) e usino la conoscenza inclusa nella stazionarietà/non stazionarietà di un fenomeno.

Pertanto, nell'ambito della ricerca di dottorato, si intendono perseguire i seguenti obiettivi:

1. esplorare i principali risultati teorici, algoritmici e sperimentali dello stato dell'arte della ricerca; individuare punti di forza e i limiti delle soluzioni proposte ed identificare estensioni e soluzioni che mirino alla considerazione della correlazione e stazionarietà dei dati;
2. estendere e/o definire nuovi modelli di rappresentazione per i dati spazio-temporali rinvenuti da stazioni di misurazione che nel tempo possono passare dallo stato attivo a non-attivo (e viceversa);
3. estendere e/o sintetizzare nuove tecniche che affrontino compiti di scoperta di conoscenza (nello specifico compiti di previsione) da serie temporali geo-riferite. Si intendono studiare soluzioni di analisi dati in grado di processare correlazione spaziale/temporale, trend, stagionalità, stazionarietà e non, cambiamento in dati geo-riferiti misurati periodicamente;
4. valutare i modelli scoperti in applicazioni di interesse concreto.

4) Motivazioni della ricerca

La globalizzazione ha accelerato in modo significativo la comunicazione e lo scambio di esperienze, ma ha anche moltiplicato la mole di dati collezionati in seguito al monitoraggio ubiquo di fenomeni economici, sociali, ecologici e/o ambientali. In tali circostanze, è pressante la richiesta di strumenti in grado di analizzare in tempo reale dati rappresentativi del comportamento di fenomeni in osservazione al fine di trarre conoscenza da tali dati utile alla previsione del comportamento futuro dei fenomeni.

Una previsione accurata del fenomeno non ancora osservato permette di anticipare azioni legate al realizzarsi di un determinato comportamento (per esempio prevedendo l'andamento dei mercati è possibile progettare opportune attività di investimento).

Vista l'indiscussa ubiquità dei sistemi di acquisizione dati, è necessario che gli strumenti di analisi dei dati siano in grado di modellare comportamenti che variano nel tempo e nello spazio, nello specifico serie-temporali (uni-dimensionali o multi-dimensionali) geograficamente distribuite.

Sia in *statistica* sia in *data mining* lo studio di serie temporali come anche l'analisi di dati spazialmente distribuiti ha una lunga tradizione. In letteratura sono descritte numerose tecniche che affrontano compiti di previsione in serie temporali come anche interpolazione/regressione in dati spaziali. Tuttavia, la maggioranza di queste tecniche considerano lo spazio o il tempo, raramente spazio e tempo contemporaneamente, e quando lo fanno difettano di scalabilità ed efficienza. Tuttavia, nelle applicazioni odierne, dove grandi volumi di dati geo-riferiti sono collezionati continuamente e velocemente (spesso a distanza di pochi minuti), l'efficienza e la scalabilità della computazione non sono meno rilevanti della accuratezza del modello. D'altra parte è ormai consolidato il punto di vista secondo il quale la qualità dei modelli (predittivi) scoperti trae vantaggio dalla considerazione della eventuale *correlazione* osservata nei dati; correlazione che può diramarsi lungo la direzione spaziale e/o temporale. In aggiunta, la computazione di un modello da serie temporali geo-riferite non può prescindere dalla considerazione della *non stazionarietà* dei dati che cambiano tanto nello spazio quanto nel tempo.

La principale motivazione della ricerca proposta rinviene dalla esigenza di nuove strategie di analisi per serie temporali geo-riferite, che siano in grado di trarre conoscenza dalla esistenza di forme correlazioni, trend e (non)stazionarietà in dati osservabili nel tempo e spazio, usare questa conoscenza per realizzare previsioni, ricercare e correggere anomalie ed eventualmente adattare questa conoscenza alla acquisizione di nuovi dati.

5) Stato dell'arte

La comprensione delle dinamiche insite nei dati rivenienti dal monitoraggio di fenomeni sociali ed economici (indicatori del mercato o della borsa, indicatori di povertà, disoccupazione), meteorologici (temperatura, umidità, pressione) o ecologici (indicatori di contaminazione), energetici (produzione di energia, irraggiamento, velocità del vento) implica la capacità di analizzare dati che variano sia nello spazio sia nel tempo. L'analisi di tali dati a scopo predittivo affonda le sue origini nell'analisi di serie temporali, l'analisi di dati geo-riferiti e l'analisi di dati spazio-temporali con specifico riferimento, per questo progetto di ricerca, a compiti di predizione numerica.

Analisi di serie temporali

Una serie temporale è una sequenza di osservazioni di una dimensione collezionata nel tempo [5]. Le osservazioni corrispondono ai valori osservati per la dimensione di misura ordinati cronologicamente e rappresentati come una sequenza di "n" coppie di misure valore-tempo,

$$Q = \{(q_1, t_1), (q_2, t_2), \ldots, (q_n, t_n)\}$$

dove (per $\boldsymbol{i = 1,2, \ldots, n}$) $\boldsymbol{q_i}$ è valore della variabile osservata, $\boldsymbol{t_i}$ è il tempo della

misura/osseervazione.

Le tecniche per l'analisi di serie temporali sono classificate in base alla natura del dato temporale (serie temporale numerica e simbolica) [26]. In questo progetto si focalizza l'attenzione sul numerico.

Le principali tecniche per l'analisi di serie temporali numeriche a scopo predittivo [5], [20], [21] assumono la possibilità di prevedere dati del futuro sulla base di regolarità osservate nel passato. In generale un modello preddittivo temporale è nella forma:

$$f(t+1) = F(y_t, y_{t-1}, \ldots, y_{t-k+1})$$

Diverse tecniche sono definite al fine di apprende ***F()*** assumendo che la serie temporale comprenda tre componenti: *tendenza*, che descrive l'andamento medio di una serie storica nel tempo e può essere crescente, decrescente o costante; *stagionalità*, che deriva dalle fluttuazioni ondulatorie di periodicità regolare e di breve periodo che si manifestano nei valori di una serie storica; *fluttuazione casuale,* che rappresenta tutte le variazioni presenti nei dati che non possono essere spiegate mediante le altre componenti.
Seguendo tale idea, si può procedere alla scomposizione di una serie temporale nelle tre componenti di tendenza, di stagionalità e di fluttuazione casuale e combinazione delle stesse in modalità additiva o moltiplicativa. In particolare, i *modelli di smoothing (ammortamento) esponenziale* [7], [22] si distinguono per la loro interpretazione di tale approccio in una maniera che risulta versatile, accurata e semplice. I modelli di smoothing sono: esponenziale semplice (modello di Brown); esponenziale con correzione di tendenza (modello di Holt) ed esponenziale con correzione di tendenza e stagionalità (modello di Winters).

In alternativa, i *modelli autoregressivi* si basano sull'idea di identificare i legami tra le osservazioni di una serie temporale in corrispondenza di diversi periodi attraverso lo studio dell'autocorrelazione tra osservazioni separate da un intervallo temporale fisso denominato latenza (lag). ARMA [3], [15], [24] è un modello autoregressivo a media mobile che rappresenta un modello misto costruito dalla combinazione di elementi autoregressivi di ordine ***p*** (che si propone di ricavare un legame di regressione lineare tra la serie storica originaria e le serie storiche ottenute per differenziazione fino all'ordine ***p***) e elementi a media mobile di ordine ***q*** (che si propone di ricavare un legame di regressione lineare tra la serie storica originaria e le serie storiche degli errori nei periodi precedenti).
Il limite del modello autoregressivo ARMA, come anche dei più semplici modelli di smoothing esponenziale, è che essi appaiono robusti solo in caso di stazionarietà dei dati.
Una serie temporale è considerata stazionaria se media e varianza non sono funzioni del tempo, mentre l'autocorrelazione è solo funzione della distanza nel tempo tra due variabili casuali coinvolte. Test per valutare la stazionarietà di un processo sono descritti in statistica (per esempio test KPSS, che prende il nome dagli autori Kwiatkowski, Phillips, Schmidt, Shin o test di Dickey e Fuller) [6], [14], [36].
In assenza di stazionarietà, si usa invece il modello autoregressivo integrato a media mobile ARIMA [3], [15], [24] che può usare il modello ARMA sulla serie temporale ottenuta mediante ***d*** differenziazioni successive della serie storica originaria. Il modello ARIMA è stato utilizzato per previsioni accurate a breve termine. Una delle sue caratteristiche è la assenza di variabile predittore (o indipendente). Il modello ARIMAX [30]. invece sostituisce ARIMA qualora i dati da processare rivengano da serie temporali

multidimensionali dove si distingue tra variabili predittore e una variabile risposta. Al fine di considerare le variabili predittore ARIMAX combina la regressione lineare e il modello ARIMA in unico processo che ben si adatta al caso di pattern stagionali che cambiano nel tempo.

Sempre in riferimento a serie temporali multidimensionali è definito il *modello di regressione con correlazione di serie* [13] che modella la correlazione temporale dei residui utilizzando un coefficiente di autocorrelazione $\boldsymbol{\rho}$:

$$y_t = \beta_0 + \sum_i x_{i,t}\beta_i + u_i$$

$$u_t = \rho \cdot u_{t-1} + \varepsilon_t$$

Una procedura iterativa per la stima di $\boldsymbol{\rho}$ è definita in [13].

Infine è rilevante per l'analisi di serie temporali l'*analisi spettrale* (tipicamente utilizzata nel campo economico) in quanto essa fornisce gli strumenti per procedere alla analisi dello spettro di frequenza di una serie temporale stazionaria o resa tale. L'esame dell'andamento dello spettro di frequenza di una serie temporale, e specialmente dei suoi picchi (massimi relativi), permette di identificare le più importanti componenti oscillatorie presenti nella serie storica e comportamenti periodici all'interno di serie temporali. Tale informazione può essere utile per decidere il modello di previsione da apprendere e la estensione della finestra di osservazioni.

Analisi di dati georeferenziati

Un collezione di dati geo-referiti [32], [35], [43] include misure di una dimensione collezionate simultaneamente, ma localizzate in diversi locazioni della superficie terrestre.

La correlazione (per misurazioni spazialmente distribuite della stessa dimensione di analisi o anche di diverse dimensioni di analisi) ha un ruolo centrale nello studio di tecniche di regressione (predire una variabile risposta in base ai valori di variabile predittore) e interpolazione (stimare una variabile risposta sulla base di altri valori osservati per la medesima variabile risposta) di dati spaziali. Il concetto di correlazione spaziale è ben sintetizzata dalla *legge della Geografia di Tobler* [41] in accordo alla quale "qualsiasi misura è correlata con qualsiasi altra misura, ma le misure più vicine sono più correlate di quelle distanti". Specificatamente per la autocorrelazione spaziale (correlazione del valore di una variabile misurata ad una locazione dello spazio con valori della medesima variabile e ad altre locazioni dello spazio) sono definiti in letteratura indicatori (indice di Moran, indice di Geary, misura di Getis and Ord) che consentono di verificarne l'esistenza e la forza sia a livello globale sia a livello locale [3], [36].

Il *k-Nearest Neighbour* è una tecnica di regressione definita in *data mining* che in maniera naturale permette di basare la interpolazione di una variabile sulla considerazione della sua attesa autocorrelazione spaziale con i valori della variabile osservati nel vicinato. Il valore sconosciuto di una risposta a una data locazione dello spazio è predetto come la media (pesata) delle risposte osservate nel vicinato. I pesi sono inversamente proporzionali alla distanza del vicino (*inverse distance weighting*

[34]), in tal modo vicini prossimi influiscono sulla predizione più di vicini distanti. La formula generale di interpolazione è la media pesata:

$$Z(s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i)$$

$\boldsymbol{Z(s_o)}$ è il valore della variabile $\boldsymbol{Z}$ da stimare nel punto $\boldsymbol{s_o}$, mentre $\boldsymbol{i = 1, 2, 3, ..., n}$ è il numero di punti noti utilizzati nella interpolazione. $\boldsymbol{\lambda_t}$ rappresenta il peso assegnato a ciascuna misura sperimentale usata nell'interpolazione. L'espressione per determinare i pesi $\boldsymbol{\lambda_i}$ è:

$$\lambda_i = \frac{d_{io}^{-p}}{\sum_{i=1}^{n} d_{io}^{-p}}$$

dove $\boldsymbol{d_{io}}$ è la distanza geografica tra il punto in cui interpolare la variabile ($\boldsymbol{s_o}$) ed il punto $\boldsymbol{s_i}$ di cui si conosce la misura della medesima variabile. $\boldsymbol{p}$ è il fattore che riduce i pesi aumentando la distanza $\boldsymbol{s_o}$.

In alternative, il Kriging [11], [36] denota una famiglia di tecniche di interpolazione spaziale che interpola una valore sconosciuto come una combinazione lineare delle osservazioni nel vicinato i cui coefficienti sono modellati tramite un modello di secondo ordine della variabile (variogramma)

$$Z(s_0) = \sum_{i=1}^{n} \lambda_i(s_0) Z(s_i)$$

I pesi $\boldsymbol{\lambda_i()}$ sono ottenuti come soluzione di un sistema di equazioni lineari per minimizzare la varianza dell'errore di predizione. Diversamente dallo IDW, dove il calcolo dei pesi è semplicemente basato sull'inverso della distanza, il Kriging basa la stima dei pesi sulla computazione del variogramma. Formalmente il variogramma approssima la dissimilarità statistica di una misura attraverso lo spazio. Più alto è il valore stimato dal variogramma, più sono diversi i valori.

La stima del variogramma ha complessità cubica rispetto alla numerosità dei dati osservati. Nel confronto tra IDW e Kriging, lo IDW è più semplice e efficiente, il Kriging è più accurato premessa una accurata stima del variogramma.

La Regressione Geografica Pesata (*Geographically Weighted Regression* (GWR)) [36] o il modello bayesiano proposto da LeSage [23] sono invece modelli di regressione spaziali definiti per dati multi-dimensionali (la variabile risposta deve essere localmente predetta sulla base dei valori delle variabili predittore). In particolare, GWR ricorre a una combinazione lineare delle variabili predittore per stimare la variabile risposta. Diversamente dalla regressione classica i coefficienti della combinazione lineare non sono globali per l'intera area di dati, ma sono stimati localmente a ciascuna locazione dello spazio in accordo alla possibile non-stazionarietà spaziale dei dati.

I modelli di auto-regressione spaziale [2] operano invece in due fasi. Nella prima fase la variabile risposta è considerata non-spaziale e per essa si apprende una tradizionale combinazione lineare delle variabili predittore quale modello di predizione. Nella seconda fase si presume una correlazione spaziale dei residui della regressione e si

modella tale dipendenza ricorrendo a una matrice "spaziale" in un approccio auto-regressivo.

$$y = x\beta + u$$

$$u = Wu + \varepsilon$$

$\boldsymbol{\beta}$ è il vettore dei coefficienti di regressione, $\boldsymbol{u}$ è il vettore dei residui, $\boldsymbol{W}$ è la matrice quadrata dei pesi spaziali aventi zeri sulla diagonale principale ($\boldsymbol{w_{ij}}$ è il peso spazialmente definito tra l'esempio $\boldsymbol{x_i}$ e l'esempio $\boldsymbol{x_j}$). Di conseguenza il modello:

$$y = x\beta + W\left(y - x\beta\right) + \varepsilon$$

i cui parametri son stimati con una generalizzazione del metodo dei minimi quadrati o con le tecniche di massima verosimiglianza.

Una evoluzione del precedente modello riportata in [8], [17] assume la variabile di risposta correlata spazialmente e dipendente dai valori di attributi nei punti vicini:

$$y = x\beta + WX\gamma + \rho Wy + \varepsilon$$

Analisi dei dati spazio-temporali

Il *data mining* spazio temporale (STDM) è un'area di ricerca relativamente recente che adatta o definisce tecniche e metodi di *data mining* classico a dati spazio-temporali. Il crescente interesse verso questo campo di ricerca ha portato alla realizzazione di diversi sistemi per l'analisi di dati spazio-temporali. Per esempio, presso il centro di ricerca UCLA (*University of California, Los Angeles*) è stato sviluppato un ambiente per l'analisi e comprensione dei dati spazio-temporali chiamato CONQUEST (*CONtext-based Querying in Space and Time*)[37]. La conferma di questo crescente interesse rinviene da numerose conferenze e workshop nel campo del *data mining spazio-temporale*.

È possibile identificare due direzioni di ricerca con riferimento al *data mining* dai dati spazio-temporali:

a) l'introduzione di elementi per la gestione del tempo nelle tecniche spaziali;
b) l'introduzione di elementi per la gestione dello spazio nelle tecniche di analisi di serie temporali.

Una collezione di dati spazio-temporale è una collezione di serie temporali dove ogni serie contiene misure successive della caratteristica in esame collezionate ad una posizione della spazio. Diverse tecniche di *data mining* spazio-temporale [31], [34], [38], [40], [44] sono state studiate per diversi compiti e impiegate in campi quali epidemiologia, ecologia, climatologia o statistica, in cui i dataset di natura spazio-temporale sono generalmente raccolti.

In particolare, il compito predittivo formulato per un dominio spazio-temporale ha recentemente attirato l'attenzione di numerosi ricercatori [12],[29].

I modelli di regressione spazio-temporale (per esempio [9], [10]) estendono la considerazione della correlazione tra i dati dalla dimensione spaziale a quella temporale. Lo studio di processi spazio-temporali si fonda quindi sulla stima della funzione di correlazione o covarianza spazio-temporale. Mentre originariamente correlazione spaziale e correlazione temporale erano studiate separatamente per poi essere combinate in maniera additiva o moltiplicativa, oggi ha preso piede l'idea di

esplorare la correlazione (o autocorrelazione) simultaneamente in tempo e spazio. Seguendo tale approccio sono stati definiti nuovi indici di misura della autocorrelazione spazio-temporale (per esempio indice di autocorrelazione spazio-temporale di Moran [3],[36]).

In particolare, la regressione spazio-temporale è essenzialmente demandata a generalizzazioni temporali del modello di regressione spaziale autoregressivo nelle quali si procede a considerare la correlazione spazio-temporale dei residui. La matrice $\boldsymbol{W}$ è in questo caso stimata usando variogrammi spazio-temporali. In maniera simile, combinazioni del modello spazio-regressivo e auto-regressivo ($y = x\beta + WX\gamma + \rho Wy + \varepsilon$) sono state già messe a punto [29]. $\boldsymbol{W}$ è inteso come un prodotto delle matrici $\boldsymbol{S}$ e $\boldsymbol{T}$ legate rispettivamente allo spazio e al tempo. Ogni esempio è considerato dipendente da un numero fisso di vicini spaziali (indipendentemente dal tempo) e da un numero fisso di vicini che lo precedono nel tempo. Il problema principale applicabile a questo modello è la stima corretta di $\boldsymbol{T}$ e $\boldsymbol{S}$.

Più recentemente [28] hanno investigato la sintesi di variabili di aggregazione in un vicinato spazio temporale che catturano fenomeni quali trend di variazione, autocorrelazione e variabilità e usato le stesse nell'apprendimento di tradizionali modelli di regressione (reti neurali, foreste di alberi di modelli) come anche modelli temporali (ARMA) impiegati nella previsione di uno specifico fenomeno spazio temporale quale le variazioni di velocità del vento in spazio e tempo.

In [16], [19], [27], [39] modelli dinamici sono stati studiati al fine di modellare nella funzione di predizione tanto la variabile predittore con un spostamento nel tempo e/o nello spazio quanto le variabili dipendenti con uno spostamento nel tempo. Tale approccio apre interessanti prospettive di ricerca legate alla gestione della alta intercorrelazione tra le variabili ritardate (multicollinearità); autocorrelazione di variabili casuali e (non)stazionarietà delle variabili modellate. L'effetto della autocorrelazione nei modelli dinamici è la stima inefficiente del metodo dei minimi quadrati a causa della violazione dell'ipotesi di indipendenza delle variabili. Tale problema è risolto in letteratura con diversi approcci che includo la riformulazione del modello al fine di migliorare la sua struttura dinamica, l'introduzione di variabili strumentali, l'applicazione del metodo dei minimi quadrati generalizzati.

6) Approccio al problema

La dottoranda intende investigare inizialmente il problema dell'apprendimento di modelli di previsione dinamici in serie temporali mono-dimensionali geo-riferite, per poi estendere eventualmente la ricerca alla considerazione di serie temporali multi-dimensionale.

In particolare, la dottoranda pianifica di esplorare la possibilità di estendere le tecniche di analisi di serie temporali tradizionali con la esplicita considerazione del fenomeno della autocorrelazione spaziale che caratterizza una dimensione geo-fisica qualora venga misurata in un vicinato spaziale. Al riguardo intende investigare tecniche di clustering spaziale già esistenti (per esempio dbscan) ed eventualmente sintetizzarne di nuove (basate sulla valutazione di indicatori di autocorrelazione locale e/o globale) al fine di individuare i confini delle regioni attraversate da istanze di dati autocorrelati spazialmente che appartengono a serie temporali.

Una regionalizzazione che si ripropone lungo l'asse temporale, impone un naturale raggruppamento delle serie temporali che sono ripetutamente raggruppate in una qualche regione ad un dato istante di tempo. Tale raggruppamento può essere inteso come la manifestazione di istanze di un medesimo comportamento. Le tecniche per la analisi di serie temporali possono essere quindi utilizzate per modellare tale comportamento piuttosto che ogni singola serie. Il modello appreso va associato a tutte le istanze del raggruppamento. Si prevede quindi di investigare tecniche di interpolazione spaziale (per esempio IDW o kriging) al fine di combinare spazialmente le previsioni che rinvengono da modelli appresi per diversi raggruppamenti. Si intende anche investigare un approccio di inferenza spaziale collettiva al fine di migliorare la accuratezza della predizione. Il risultato atteso sarà frutto di una soluzione algoritmica che processa sia la dimensione spaziale sia la dimensione temporale e che si presuppone debba dimostrarsi più robusta della soluzione algoritmica tradizionale che apprende un modello temporale per ciascuna stazione o un modello spaziale per ciascuna fotografia dei dati a un fissato istante di tempo.

I temi di ricerca centrali da affrontare nella realizzazione del progetto sono descritti nel seguito:

1. Sintesi di algoritmi per la *regionalizzazione* dei dati al fine di tracciare i confini spaziali della autocorrelazione di un dato (ed eventualmente la correlazione tra dati misurati per diverse dimensioni); valutazione delle possibile *evoluzioni di tale modello regionale* nel tempo: approccio statico (apprendere una regionalizzazione per ogni istante di tempo e poi individuare regolarità nelle regionalizzazioni acquisite lungo una finestra temporale) vs approccio dinamico (adattare incrementalmente la regionalizzazione appresa nel passato ai dati correnti).
2. Analisi degli *indicatori di autocorrelazione spaziale* (globale vs locale) e loro utilizzo nella regionalizzazione dei dati e nella sintesi di nuove variabili di aggregazione usate nell' apprendimento dei modelli di previsione; studio di una eventuale estensione della definizione propriamente spaziale degli indicatori di autocorrelazione verso una definizione spazio- temporale.
3. Analisi degli algoritmi di *previsione per serie temporali*, scelta della finestra temporale, studio di test di stazionarietà al fine di indirizzare la scelta dell'algoritmo di modellazione della serie temporale da utilizzare; studio degli algoritmi di *interpolazione e/o regressione spaziale* definiti al fine di migliorare l'accuratezza dei tradizionali algoritmi di analisi di serie temporale.
4. Applicazione reale (previsione e scoperta di anomalie) delle soluzioni investigate: analisi di serie temporale in *climatologia e/o ecologia*, *pianificazione energetica*, *analisi finanziaria*. Apprendimento dei modelli previsione sulla base delle metodologie acquisite; analisi comparativa con soluzioni già esistenti in termini di accuratezza e efficienza.

7) Ricadute applicative

Si presentano alcuni alcune applicazioni in cui possono essere utilizzate le tecniche di predizione sintetizzate per serie temporali geo-riferite.

1. Previsione del comportamento di fenomeni geo-fisici a breve e lungo termine al fine di anticipare e pianificare attività e comportamenti. Tale compito ha molteplici applicazioni in ambito finanziario, ambientale e socio-economico

(per esempio prevedere flessioni del mercato, intensità del traffico, condizioni meteorologiche);
2. Scoperta di anomalie (confrontare il valore predetto con il valore reale al fine di segnalare anomalie e/o cambiamenti nei dati).

8) Fasi del progetto

Nel primo anno della scuola di Dottorato in Informatica, il lavoro del dottorando sarà articolato in tre fasi:

1A) Studio approfondito della letteratura relativa all'analisi di tecniche di data mining (predittivo e non) per dati spazio-temporali. In particolare si investigheranno gli strumenti per la misura della correlazione spaziale e spazio-temporale, tecniche di previsione a breve e lungo termine data processi stazionari e non, test di stazionarietà, modelli di regressioni e interpolazione spaziale, soluzioni per l'apprendimento di modelli di predizione spazio-temporale. Si analizzeranno punti di forza e punti di debolezza delle soluzioni computazionali delineate in letteratura.

1B) Frequenza dei corsi/seminari organizzati dalla scuola ed espletamento dei relativi esami.

1C) Partecipazione a scuole internazionali, conferenze, seminari su temi di ricerca riguardanti il data mining spazio-temporale.

Nel secondo anno, il lavoro del dottorando sarà articolato in cinque fasi:

2A) Sintesi, progettazione e realizzazione di uno o più algoritmi di data mining spazio-temporale per l'analisi a scopo predittivo di serie temporali geo-riferite.

2B) Test preliminari delle soluzioni algoritmiche con dataset artificiali e/o reali.

2C) Frequenza dei corsi/seminari organizzati dalla scuola ed espletamento dei relativi esami.

2D) Collaborazione con gruppi di ricerca nazionali e/o internazionali che siano di rilevanti per l'attività di ricerca condotta con tale progetto.

2E) Pubblicazione dei risultati della ricerca condotta in riviste e conferenze di carattere internazionale e nazionale.

Nel terzo anno, il lavoro del dottorando sarà articolato in cinque fasi:

3A) Acquisizione di dati reali e valutazione sistematica delle soluzioni algoritmiche proposte in uno o più contesti applicativi.

3B) Pubblicazione dei risultati empirici in riviste/conferenze di prestigio internazionale e nazionale.

3C) Esplorazione e interpretazione dei risultati empirici e/o teorici ottenuti tramite lo svolgimento delle attività svolte negli anni precedenti.

3D) Delineazione di eventuali sviluppi futuri del lavoro.

3E) Stesura del manoscritto di tesi.

Per la valutazione empirica si intende utilizzare i dataset descritti in Sezione 10 ed eventualmente individuare nuove fonti di dati artificiali e/o reali. Si intende valutare i

risultati sulla base delle metriche definite dalla letteratura di riferimento e mostrare tramite confronto che la nuove soluzioni migliorano (statisticamente) soluzioni già definite in letteratura.

Attività	Anno I				Anno II				Anno III			
	Trim. I	Trim. II	Trim. III	Trim. IV	Trim. I	Trim. II	Trim. III	Trim. IV	Trim. I	Trim. II	Trim. III	Trim. IV
1A												
1B												
1C												
2A												
2B												
2C												
2D												
2E												
3A												
3B												
3C												
3D												
3E												

9) Valutazione dei risultati

Nella valutazione delle soluzioni algoritmiche sintetizzate si intende considerare alcuni dei contesti applicativi tra quelli descritti in Sezione 7.

Al momento sono stati reperiti dati serie temporali geo-riferite relative alla meteorologia:

1. temperatura collezionata mensilmente da 6477 stazioni meteo dislocate o simulate (per interpolazione) in Sud America dal 1960 al 1990 (http://climate.geog.udel.edu/ climate/html pages/archive.html);
2. precipitazione e temperatura collezionate mensilmente da 20590 stazioni per le precipitazione e 7280 stazioni per la temperature distribuite sulla superficie terrestre e operanti dal 1890 al 1999 (ftp://ftp.ncdc.noaa. gov/pub/data/ghcn/v2/);
3. temperatura, umidità luminosità e voltaggio misurati ogni 31 secondi da 54 sensori distribuite nel Intel Berkeley Laboratory (http://db.csail.mit. edu/labdata/labdata.html).
4. le misurazioni della velocità del vento di 1326 stazioni ad 80 metri di altezza nella regione orientale degli Stati Uniti raccolti ad intervalli di 10 minuti durante l'anno 2004 pubblicati da DOE/NREL/ALLIANCE (http://www.nrel. gov/).

10) Eventuali referenti esterni al Dipartimento

Luis Torgo, Department of Computer Science, University of Porto.

11) Riferimenti bibliografici

[1] Anselin, L. *Spatial econometrics.* In a companion to theoretical econometrics, ed. Baltagi, Oxford: Basil Blackwell, (2001a), 310-330.

[2] Anselin, L. *Spatial Econometrics: Methods and Models*. Dordrecht: Kluwer, 1988.

[3] Astutik, S., Rahayudi, B., Iskandar, A., Fitriani, R., Murtini, S. *Detection of Spatial-Temporal Autocorrelation using Multivariate Moran and Lisa Method on Dengue Hemorrhagic Fever (DHF) Incidence*. East Java, Indonesia, European Journal of Scientific Research ISSN 1450-216X Vol.49 No.2 (2011), pp. 279-285.

[4] Baltagi, B.H., Song, S.H., Koh, Y. *Testing panel data regression models with spatial error correlation*. Journal of Econometrics, 117 (1) (Nov. 2003), 123-150.

[5] Box, G., Jenkins, G., Reinsel, G. *Time Series Analysis.* Forecasting and Control, Third ed., Prentice Hall, Upper Saddle Hill, NJ, 1994.

[6] Brocklebank, J.C., Dickey, D.A. *SAS System for Forecasting Time Series.* 1986 Edition, Cary, North Carolina: SAS Institute Inc, 1986.

[7] Brown, R. G. *Exponential Smoothing for Predicting Demand*. Cambridge, Massachusetts: Arthur D. Little Inc. (1956), pp. 15.

[8] Burridge, P. *Testing for a common factor in a spatial auto-regression model*. Environment and Planning A, 13 (1981), pp. 795-800.

[9] Chasco, C., Lopez, F.A. *Space –time lags: specification strategy in spatial regression models*. First seminar of spatial econometrics Jean Paelinck. 2004.

[10] Chasco, C., Lopez, F.A. *Time-trend in spatial dependence: specification strategy in the first-order spatial autoregressive model.* Estudios de Economia Aplicada, vol.25-2, 2007.

[11] Cressie, N. *Statistics for spatial data*. Wiley, New York, 1993.

[12] Cressie, N., Majure, J.J. *Spatio-temporal statistical modeling of livestock waste in streams*. Journal of Agricultural, Biological, and Environmental Statistics 2 (1997), pp. 24-47.

[13] Davidson, R., Mc Kinon, J. *Estimation and Inference in Econometrics*. Oxford Univ.Press., New York, 1993.

[14] Dickey, D. A., Fuller, W.A. *Distribution of the Estimators for Autoregressive Time Series With a Unit Root*, Journal of the American Statistical Association 74 (366), (1979), 427-431.

[15] Donald B.P., Walden, A.T. *Spectral Analysis for Physical Applications.* Cambridge University Press, 1993.

[16] Elhorst, J.P. *Dynamic models in space and time*. Geographical Analysis 33 (2001), 119-140.

[17] Florax, R., Folmer, H. *Specification and estimation of spatial linear regression models-Monte Carlo evaluation of pre-test parameters*. Regional Science and Urban Economics, 22 (1992), pp. 405-432.

[18] Geweke, J. *The dynamic factor analysis of economic time-series models.* in D. Aigner & A. Goldberger, eds, Latent Variables in Socio-Economic Models, North-Holland, New York, 1977.

[19] Granger, C., Engle, R. *Dynamic model specification with equilibrium constraints: Co-integration and error-correction.* Discussion Paper 85-18, University of California, San Diego, 1985.

[20] Hamilton, J.D. *Time series analysis*, Princeton, N.J. 1994.

[21] Harvey, A.C. *Time series models*, Hemel Hampstead, 1993.

[22] Holt, C. C. *Forecasting Trends and Seasonal by Exponentially Weighted Averages*. *Office of Naval Research Memorandum* **52**. reprinted in Holt, Charles C. (January–March 2004)

[23] LeSage JP. *A family of Geographically Weighted Regression Models*. in Anselin, L., Florax RJGM. Rey SJ. (eds) *Advance in Spatial Econometrics*, Springer, Berlin (2004) 241-266.

[24] Mills, T. C. *Time Series Techniques for Economists.* Cambridge University Press, 1990.

[25] Molenaar, P. *A. Dynamic factor model for the analysis of multivariate time series.* Psychometrika 50 (1985), pp. 181–202.

[26] Mörchen, F. *Time Series Knowledge Mining*. Ph.D. Thesis, Philipps University, Marburg, Germany from:www.mybytes.de/papers/moerchen06tskm.pdf, 2006.

[27] Nogales, F., Contreras*, J.,* Conejo*, A.,* Espínola, R. *Forecasting Next-Day Electricity Prices by Time Series Models.* IEEE transactions on power systems, vol. 17, NO. 2 (2002), pp. 342-347.

[28] Oshashi, O., Torgo, L. *Wind speed forecasting using spatio-temporal indicators*, Open Access by IOS Press, 10.3233/978-1-61499-098-7-975

[29] Pace Kelly, R., Gilley, O. *Generalizing the OLS and grid estimators*. Real Estate Economics, 26 (1998), pp. 331-347.

[30] Pankratz, A. *Forecasting with Univariate Box-Jenkins Models: Concepts and Cases*. New York: John Wiley & Sons, Inc, 1983.

[31] Roddick, J., Hornsby, K. *Temporal, Spatial, and Spatio-Temporal Data Mining*. In First Int'l Workshop on Temporal, Spatial and Spatio-Temporal Data Mining (2000)

[32] Segurado, P., Araujo, M., Kunin, W.E. *Consequences of spatial autocorrelation on niche-based models*. Journal of Applied Ecology 43 (2006), 433-444.

[33] Shekhar, S. Chawla, S. Ravada, S. Fetterer, A. Liu, X., Lu, C.T. *Spatial databases*: *Accomplishments and research needs*. IEEE Transactions on Knowledge and Data Engineering 11, 1 (1999), 45–55.

[34] Shepard, Donald, *A two-dimensional interpolation function for irregularly-spaced data*. *Proceedings of the 1968 ACM National Conference*. (1968), pp. 517–524

[35] Steinbach, M., Tan, P., Kumar, V., Potter, C., Klooster, S., Torregrosa *A. Data Mining for the Discovery of Ocean Climate Indices.* In Proc of the Fifth Workshop on Scientific Data Mining (2002)

[36] Stewart Fotheringham, A., Brunsdon, C., Charlton, M., *Geographically Weighted Regression.* The Analysis of Spatially Varying Relationships , Wiley, 2002

[37] Stolorz, P., Mesrobian, E., Muntz, R., Santos, J.R., Shek, E., Yi, J. Mechoso, C., Farrara, J. *Fast Spatio-Temporal Data Mining from Large Geophysical Datasets.* In proc. of 1st Int. Conference on Knowledge Discovery and Data Mining (August 20-21, 1995)

[38] Streitberg, B. *Multivariate Models of Dependent Spatial Data*. Statistical Analysis of Spatial Data (1979), 139-177.

[39] Takens, F. *Detecting strange attractors in turbulence*. Dynamical systems and turbulence Warwick 1980, 898(1)6–381 (1981), pp. 366-366.

[40] Tan, M., Steinbach, P., Kumar, V., Potter, C., Klooster, S., Torregrosa, A. *Finding Spatio-Temporal Patterns in Earth Science Data*. In KDD 2001 Workshop on Temporal Data Mining (2001)

[41] Tobler, W. *Mathematical Map Models.* Proceedings, International Symposium on Computer Aided Cartography, Reston, ACSM (1975), pp. 66-73.

[42] Upton, G., Fingleton, B. *Spatial Data Analysis by Example.* Vol. 1: Point Pattern and Quantitative Data, Wiley. New York, NY, 1985.
[43] Wartenberg D. *Multivariate spatial correlation: A method for exploratory geographical analysis.* Geographical Analysis 17 (1985), 263–283.
[44] Worboys, M.F. *GIS - A Computing Perspective*. Taylor and Francis, 1995.