A R E S - S P A D A

Karel BAŘINA a Radim ŠTAMPACH

Brno 2005

ARES

ARES je systém pro generování asociačních pravidel z geografických dat.

Vstupy do systému:

- prostorová databáze
- množina referenčních objektů
- množina „task-relevant“ objektů
- prostorová hierarchie (BK – background knowledge)
- počet úrovní v hierarchii prostorových objektů (vyšší číslo úrovně *l* znamená nižší úroveň hierarchie, např. republika (0), kraj (1), okres (2))
- minimální prahové hodnoty support a confidence pro každou z úrovní hierarchie *l*, (*minsup[l]* a *minconf[l])*. Cílem systému ARES je nalezení víceúrovňových asociačních pravidel

Referenční objekty jsou hlavním předmětem popisu asociačních pravidel, zatímco task-relevant objekty jsou prostorové objekty s vazbou na referenční objekty. Například můžeme hledat asociační pravidla popisující vazby mezi velkými městy (referenčními objekty) pomocí vlastností ostatních prostorových objektů (task-relevant objektů), jako vodní toky a silnice. Každou množinu task-relevant objektů si můžeme představit jako jednu vrstvu GIS obsahující vztahy mezi objekty na základě jejich geometrie. Aby bylo pracovat s více úrovněmi hierarchie prostorových objektů jednotným způsobem, objekty v těchto úrovních jsou uspořádány do jedné nebo více vrstev definovaných uživatelem tak, že četnost vzorů (frequency of patterns), stejně jako síla pravidel (strength of rules), závisí na dané úrovni definované hierarchie. Přesněji řečeno, vzor *P* se supportem *s* na úrovni hierarchie *l* je častý (frequent), jestliže *s>=minsup[l]* a všichni jeho předchůdci na vyšších úrovních *l* jsou rovněž častí. Asociační pravidlo *Q->R (s%, c%)* na úrovni *l* je silné (strong), jestliže jeho vzor *P (s%)* je častý a zároveň jeho confidence *c>=minconf[l]*.

Pozn.: asociační pravidlo: $P \rightarrow Q(s\%, c\%)$

prostorový vzor: $P \cup Q$

$P \cap Q = \{\}$

parametr *s* (support) je pravděpodobnost $p(P \cup Q)$

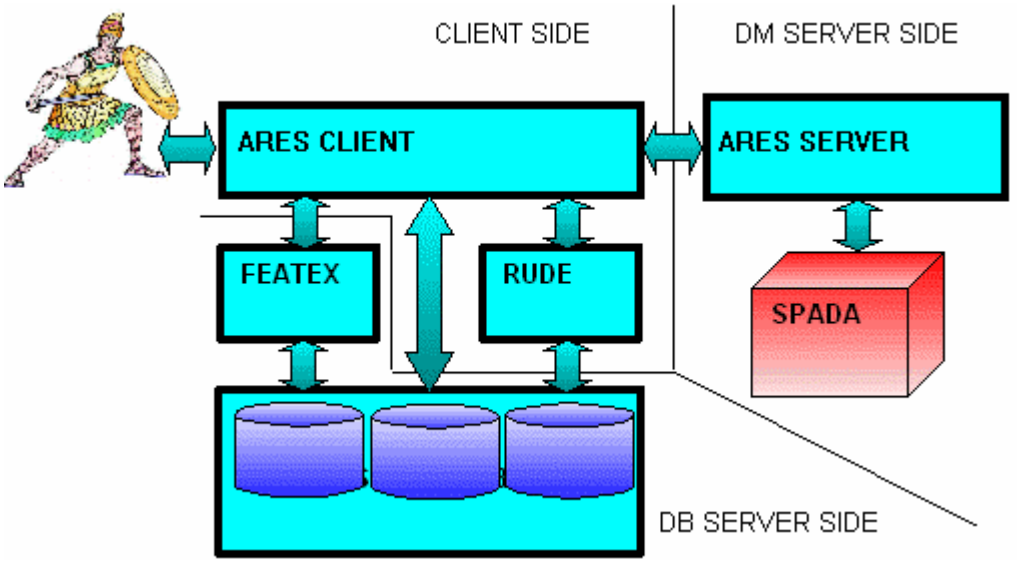
parametr *c* (confidence) je pravděpodobnost $p(Q|P)$

SPADA

Systém ARES využívá k generování asociačních pravidel algoritmus SPADA (Spatial Pattern Discovery Algorithm). Autorem algoritmu SPADA napsaného v jazyce Prolog je Donato Malerba z univerzity v Bari v Itálii. SPADA je založena na metodách induktivního logického programování a umožňuje vyhledávat pravidla ve více úrovních hierarchie (multi-level spatial association rules). Při generování pravidel využívá logiku prvního řádu.

Zajímavá asociační pravidla mohou být s větší pravděpodobností objevena na nižších úrovních hierarchie. Na druhou stranu, silnější pravidla (vysoký support) nastane s větší pravděpodobností na vyšších úrovních hierarchie.

Architektura systému ARES



Obr. 1. Architektura systému ARES

Systém ARES má tři základní části.

1. Na databázovém serveru je samotná databáze Oracle a modul FEATEX (balík funkcí a procedur, z nichž každá zjišťuje v databázi zvláštní vlastnosti (feature)):

- polohové (např. souřadnice centroidu)
- geometrické (např. velikost, obvod)
- směrové (vzájemná orientace objektů v 2D nebo 3D)

- topologické (např. vzájemné protínání, překrývání)
- hybridní (spojení dvou nebo více kategorií vlastností, např. rovnoběžnost – geometrické a zároveň topologické)

2. ARES Server je samotná implikace algoritmu SPADA vyvolatelná pomocí klienta.

3. ARES Client představuje grafické uživatelské rozhraní vytvořené v Javě pro přístup do vzdálené databáze a vyvolání algoritmu SPADA z ARES Serveru. Skládá se z GUI a modulů RUDE a WISDOM+. RUDE je doplněním FEATEXu a provádí diskretizaci numerických dat. WISDOM+ může být použit k extrahování dat např. z obrázku a jejich uskladnění v databázi, umí rozlišit text, grafiku, atd. Ve verzi, která je volně ke stažení, není WISDOM+ implementován.

Instalace systému ARES a připojení na databázový server

Součásti instalace systému ARES pro Windows:

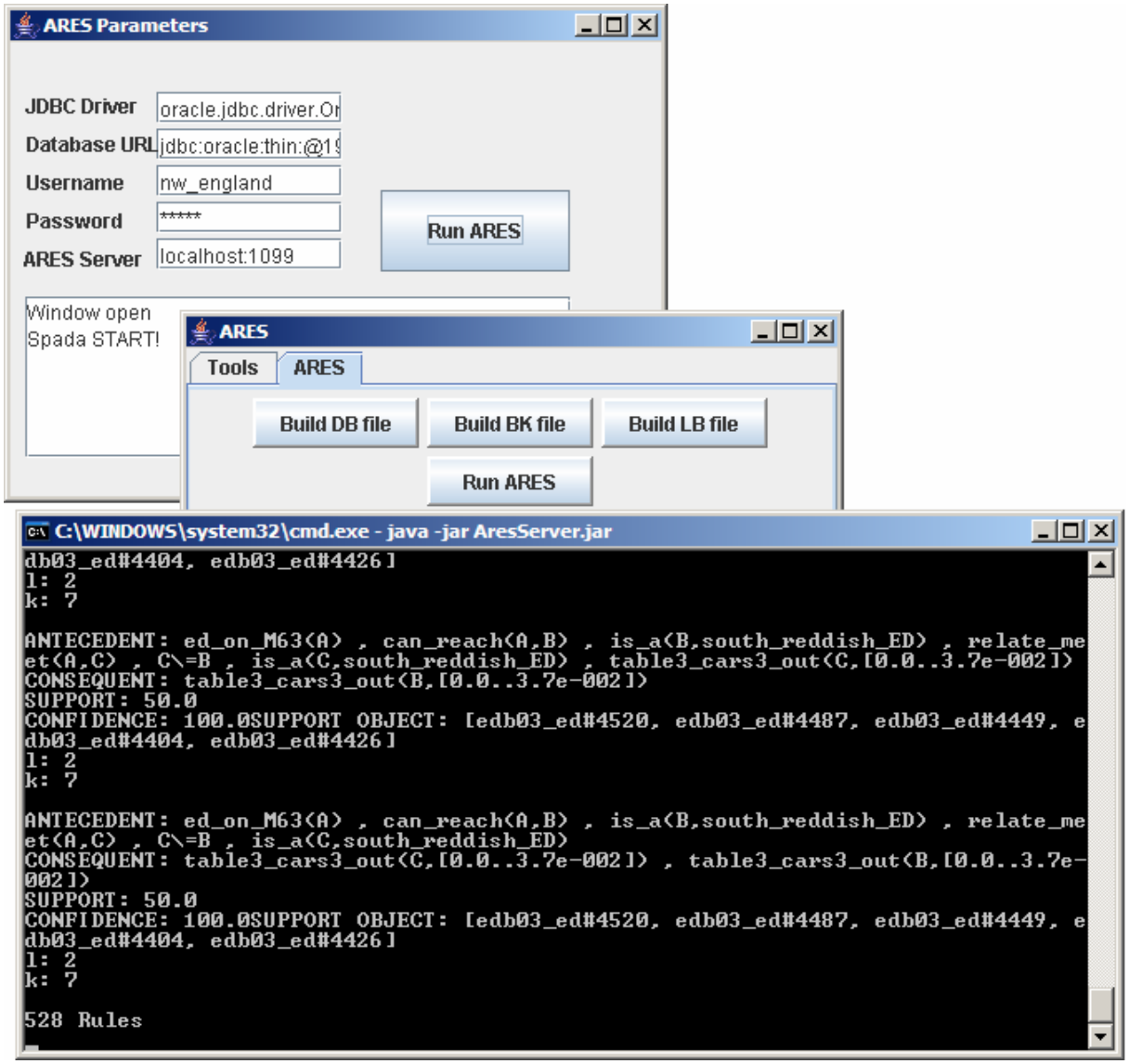
FEATEX component	http://www.di.uniba.it/~malerba/software/ARES/featex.zip
Algoritmus SPADA	http://www.di.uniba.it/~malerba/software/ARES/SpadaExe.zip
ARES Server	http://www.di.uniba.it/~malerba/software/ARES/AresServer.jar
ARES Client	http://www.di.uniba.it/~malerba/software/ARES/AresClient.jar
Autoclass library	http://www.di.uniba.it/~malerba/software/ARES/autoclass.dll

Soubor SpadaExe.zip je třeba rozbalit do adresáře C:\SPADA, kam je vhodné umístit i soubory AresServer.jar a AresClient.jar. Pro spuštění ARESu je nutné nejprve spustit v příkazové řádce ARES Server příkazem

java -jar AresServer.jar

V dalším kroku je již možné otevřít grafické rozhraní systému spuštěním souboru AresClient.jar. Objeví se okno, ve kterém je připravená adresa vzdáleného databázového serveru s Oracle, uživatelské jméno a heslo. Stačí stisknout tlačítko „Run ARES“. V novém okně je důležitá druhá záložka „ARES“ a opět tlačítko „Run ARES“. Nabídne se okno pro

vybrání vstupního souboru .db. Po úspěšném vygenerování asociačních pravidel zbývá v dalším okně zvolit umístění výstupních souborů .xml.



Obr. 2. Grafické uživatelské rozhraní systému ARES s výpisem průběhu generování asociačních pravidel

Algoritmus SPADA je možné spustit i bez grafického prostředí ARES Klienta přímo z příkazové řádky, nepotřebujeme-li pracovat přímo se vzdálenou databází Oracle. Je třeba mít opět program OOSPADA.EXE v adresáři C:\SPADA a zadat následující příkaz:

oospada.exe XXX,

kde XXX je shodný název souborů .bk, .db a .lb (bez přípony).

Vstupní a výstupní data systému ARES

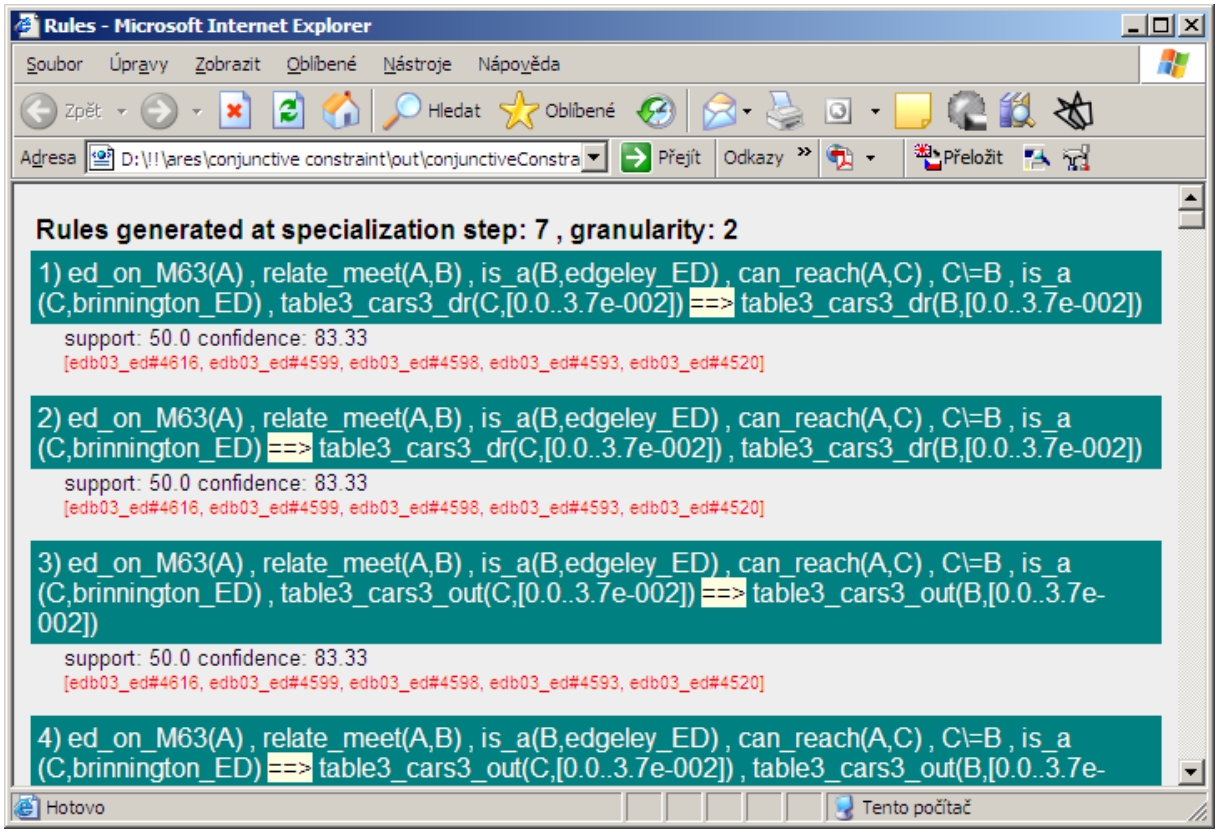
Vstupní data musí být uložena ve formě tří souborů:

1) .db: výstup modulů FEATEX a RUDE aplikovaného na data z prostorové databáze, jakmile je tento soubor vytvořen, práce s databází Oracle končí. Dále je používán pouze soubor .db.
2) .lb: language bias – omezuje vyhledávací prostor, obsahuje zadané parametry *minsup* (viz výše), *minconf* (viz výše), *max_level* (určuje, do jaké úrovně v hierarchii mají být asociační pravidla vyhledávána), *max_ref_steps* (maximální délka vzoru)
3) .bk: background knowledge (definuje hierarchii, udává vztahy nadřazenosti a podřazenosti mezi objekty) a definuje vztahy, které se objeví v generovaných pravidlech.

   Např. spojení „can_reach“ (vzájemná dosažitelnost po dálnici) je generováno pro objekty A a B, pokud FEATEX vygeneroval pro A i pro B vztah „relate_comes_from“ nebo „relate_crosses“ ve vztahu k dálnici M63 – dálnice v nich začíná nebo jimi prochází. Pak jsou po ní vzájemně dosažitelné.

Výstupní data jsou uložena ve formátu XML, jeden soubor pro každou hierarchickou úroveň. Pravidla by mělo být možné prohlížet a řadit prostřednictvím souboru HTML, který ovšem nefunguje v MS Internet Exploreru ani Mozilla Firefoxu. Samotné soubory XML je možné otevřít pomocí MS Internet Exploreru.

Pokud je algoritmus SPADA spuštěn výše popsaným způsobem bez grafického prostředí, vygeneruje pro každou úroveň hierarchie (level) a pro každý počet literálů v pravidle (steps) nejen soubor .XML se samotnými asociačními pravidly splňujícími hodnoty minsup a minconf, ale také soubory .OUT a .PAT. Soubor .OUT obsahuje všechny vygenerované vzory (patterns) a následně soupis všech vzorů, které vyhověly a těch, které nevyhověly zadané hodnotě minsup. Jsou zde také obsažena všechna vygenerovaná pravidla, která jsou rozdělena, podle toho, zda splnila či nesplnila zadanou hodnotu minconf. Soubor .PAT obsahuje opět výpis vzorů, které splnily podmínku minsup. Při spuštění algoritmu SPADA pomocí grafického rozhraní jsou tyto soubory automaticky smazány.

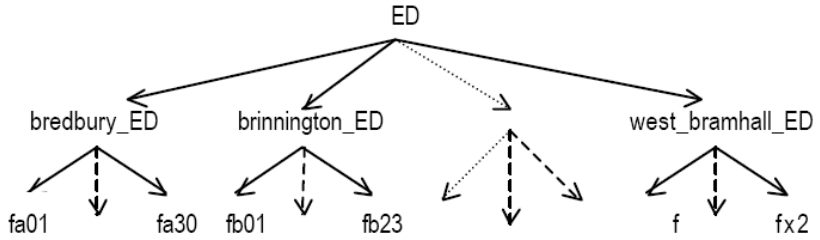


Obr. 3. Ukázka výstupního souboru XML obsahujícího výpis asociačních pravidel

Experiment

Byl proveden experiment na datech získaných z britského sčítání lidu 1991 v oblasti Stockport, což je jeden z metropolitních okresů oblasti Greater Manchester. Data jsou uložena ve vzdálené databázi Oracle, na kterou je adresa přednastavena v programu ARES Klient. Vstupní data pro systém ARES jsou dostupná na stránce s dokumentací k systému ARES (viz odkazy).

Cílem bylo zjistit, jestli je možné vygenerovat pomocí algoritmu SPADA pravidla vhodná pro dopravní plánování, především pravidla související s dojížděním v dané oblasti, kterou prochází dálnice M63 (regiony Brinnington, Cheadle, Edgeley, Heaton Mersey a South Reddish). Každý region se dále člení na sčítací okrsky (viz obr. 4). Tato hierarchie je definována v souboru .bk.



Obr. 4. Hierarchie dat (administrativní členění oblasti Stockport)

O každém sčítacím okrsku jsou v databázi uvedena následující data:

- *dr_out*: osoby dojíždějící do práce mimo svůj sčítací okrsek
- *cars3_dr*: osoby samostatně výdělečně činné a zaměstnanci vyjíždějící za prací, kteří mají v domácnosti tři a více automobily
- *cars3_out*: osoby samostatně výdělečně činné a zaměstnanci vyjíždějící za prací mimo svůj sčítací okrsek, kteří mají v domácnosti tři a více automobily

Tab. 1. Hodnoty *minsup* a *minconf* použité v experimentu pro jednotlivé hierarchické úrovně

Level	minsup	minconf
1	70%	90%
2	50%	80%
3	3%	80%

V souboru .lb lze nastavit ještě parametry *max_level* (udává, do jaké hierarchické úrovně mají být pravidla vyhledávána) a *max_ref_steps* (maximální délka vzoru). Pro experiment bylo například použito zadání pro vyhledávání pravidel bez omezení (no constraint) a také zadání omezené na ta asociační pravidla, která obsahují alespoň jednu z hodnot *dr_out*, *cars3_dr* nebo *cars3_out* (no pure spatial patterns). Rozdíl spočívá v přidání následujícího parametru do souboru .lb ve druhém případě:

pattern_constrait(dr_out(_,_),cars3_dr(_,_),cars3_out(_,_),1)

V prvním případě SPADA vygenerovala 18567 pravidel, z nichž mnohá byla pro zadané šetření nevyužitelná. Po výše uvedeném omezení bylo vygenerováno 14204 asociačních pravidel.

Pro samotný experiment byla použita data nabízená na stránkách k dokumentaci systému ARES. Ve zipovém souboru (in_noConstraint.zip) jsou vždy všechny tři potřebné vstupní

soubory. Pro kontrolu je dostupný i druhý soubor ZIP s výstupnými daty (out_noConstraint.zip).

Vstup: http://www.di.uniba.it/~malerba/software/ARES/in_noConstraint.zip
Výstup: http://www.di.uniba.it/~malerba/software/ARES/out_noConstraint.zip

No Constraint:

Bez omezení parametrem Constraint_pattern (**No Constraint**) bylo vygenerováno mnoho asociačních pravidel 18567, z nichž však bylo mnoho nic neříkajících, např.

step 3, level 2

3) ed_on_M63(A) , close_to(A,B) ==> is_a(B,edgeley_ED)
support: 100.0 confidence: 100.0

výklad: Jestliže A je blízko B, přičemž A leží na dálnici M63, pak platí při support a confidence 100, že B leží v regionu Edgeley.

Proto bylo přijato omezení, které mělo zajistit generování pravidel obsahující jak zmínku o statistických datech, tak i polohový vztah - **No Pure spatial constraint.** Do souboru .lb byl přidán následující parametr:

pattern_constraint([table3_dr_out(_,_), table3_cars3_dr(_,_), table3_cars3_out(_,_)],1).
pattern_constraint([can_reach(_,_),close_to(_,_),relate_meet(_,_)],1).

Vygenerováno bylo 18567 asociačních pravidel.

Př. relativně zajímavého pravidla: step 6, level 2

424) ed_on_M63(A) , can_reach(A,B) ==> is_a(B,heaton_mersey_ED) , table3_dr_out(B,[0.2857..0.4782]) , table3_cars3_out(A,[0.0..3.7e-002]) , table3_cars3_dr(B,[0.0..3.7e-002])
support: 90.0 confidence: 90.0

výklad: Jestliže A a B jsou vzájemně dostupné po dálnici, pak platí při support a confidence 90, že B leží v regionu Heaton Mersey, přičemž podíl vyjíždějících z B je mezi 28,57-47,82%, ale podíl dojíždějících s 3 a více auty je jen do 0,37%. Zároveň platí, že podíl těchto dojíždějících s 3 a více v A je rovněž do 0,37%.

I přes určité omezení je pravidel mnoho. Následovalo omezení, aby byla generována pravidla vždy s 2 sčítacími okrsky ED, které budou mít podobné statistiky – stejný podíl vyjíždějících apod. (**Conjuctive constraint**):

pattern_constraint([[table3_dr_out(X,Z), table3_dr_out(Y,Z),X\=Y], [table3_cars3_dr(X,Z), table3_cars3_dr(Y,Z),X\=Y], [table3_cars3_out(X,Z), table3_cars3_out(Y,Z),X\=Y]],1).
lb_required_atoms([can_reach(_,_),close_to(_,_),relate_meet(_,_)],1).

Vygenerováno bylo jen 528 pravidel.

Př. vygenerovaného pravidla při Conjuctive constraint (level 2)

176) ed_on_M63(A) ==> can_reach(A,B) , is_a(B,cheadle_ED) , can_reach(A,C) , C\=B , is_a(C,edgeley_ED) , table3_cars3_dr(C,[0.0..3.7e-002]) , table3_cars3_dr(B,[0.0..3.7e-002])
support: 100.0 confidence: 100.0

výklad: Při hodnotě support a confidence 100 platí, že B leží v regionu Cheadle, C leží v regionu Edgeley. Přitom platí, že B a C nejsou totožné, jsou obě dosažitelné po dálnici z A, a B i C mají podíl vyjíždějících z rodin s 3 a více auty do 0,37%.

Odkazy

Donato Malerba – ARES
http://www.di.uniba.it/~malerba/software/ARES

M. Berardi, A. Appice, M. Ceci, D. Malerba (2004). Mining spatial data discovery of spatial association rules with ARES, *Symbolic and Spatial Data Analysis : Mining Complex Data Structures*, 31-43.

D. Malerba, F. Esposito, F.A. Lisi & A. Appice (2002). Mining spatial association rules in census data, *Research in Official Statistics*, 5, 1, 19-44.
http://www.di.uniba.it/~malerba/software/ARES/..%5C..%5Cpublications%5CROS.pdf

A. Appice, M. Ceci, A. Lanza, F.A. Lisi, & D. Malerba (2003). Discovery of Spatial Association Rules in Georeferenced Census Data: A Relational Mining Approach, *Intelligent Data Analysis*, 7, 6.
http://www.di.uniba.it/~malerba/software/ARES/..%5C..%5Cpublications%5Cida00146.pdf

D. Malerba, F.A. Lisi, A. Appice & F. Sblendorio (2003). Mining Census and Geographic Data in Urban Planning Environments. In L. Santini& D. Zotta (Eds.), Terza Conferenza Nazionale su Informatica e Pianificazione Urbana e Territoriale, Alinea Editrice, Firenze, Italia.
http://www.di.uniba.it/~malerba/software/ARES/..%5C..%5Cpublications%5Cinput03.pdf

M. Ceci, A. Appice, & D. Malerba (2004). Spatial Associative Classification at Different Levels of Granularity: A Probabilistic Approach, in J.-F. Boulicaut, F. Esposito, F. Giannotti, & D. Pedreschi (Eds.), *Knowledge Discovery in Databases: PKDD 2004*, Lecture Notes in Artificial Intelligence, 3202, 99-111, Springer, Berlin, Germany.
http://www.di.uniba.it/~malerba/software/ARES/..%5C..%5Cpublications/PKDD04.pdf