



**Escola Politècnica Superior  
de Castelldefels**

UNIVERSITAT POLITÈCNICA DE CATALUNYA

# TREBALL DE FI DE CARRERA

**TÍTOL DEL TFC: Anàlisi Espai-Temporal Multiresolució de Matrius de Teletrànsit**

**TITULACIÓ: Enginyeria Tècnica de Telecomunicació, especialitat Telemàtica**

**AUTOR: Isaac Balasch**

**DIRECTOR: David Rincón**

**DATA: 15 de Juny del 2010**



**Títol:** Anàlisi Espai-Temporal Multiresolució de Matrius de Teletrànsit

**Autor:** Isaac Balasch

**Director:** David Rincón

**Data:** 15 de Juny del 2010

## Resum

En aquest document, es mostra la importància de les matrius de teletrànsit, les seves característiques i el mode en que són tractades a nivell matemàtic a través de models de difusió.

Les matrius de teletrànsit, corresponents al volum d'informació que es transfereix d'un node d'una xarxa a qualsevol dels altres nodes, tenen molta rellevància en l'enginyeria de xarxes, especialment pel que fa a la recerca de models predictius del comportament d'aquestes.

Els estudis realitzats amb matrius de teletrànsit, conjuntament amb models matemàtics com las wavelets de difusió, permeten caracteritzar el comportament d'una xarxa, fent prediccions sobre el seu comportament i el que és de vital importància pels operadors de les xarxes, configurar l'encaminament de forma adient.

En el document es mostraran els conceptes matemàtics més rellevants per justificar la utilització de les wavelet de difusió a més de mostrar altres operadors amb els que s'ha treballat fins el moment per comparar-los amb el desenvolupat en aquest treball.

S'ha realitzat un estudi en què s'ha aplicat un operador de difusió basat en el model de gravetat incorporant en el càlcul un valor de correlació temporal mesurat sobre les matrius.

S'ha obtingut un operador que aconsegueix millors resultats que els obtinguts fins el moment amb altres operadors de difusió. Els resultats obtinguts, amb dades reals de les xarxes Abilene (xarxa nord-americana) i Geant (xarxa acadèmica europea) s'han mostrat des de las perspectives de l'error quadràtic mig (MSE), i des de l'àmbit de la compressibilitat de l'energia de la matriu en la mínima quantitat de coeficients en el domini transformat..

S'ha obtingut uns resultats bons, tot i així hauran de ser verificats per estudis més amplis, degut a la falta de capacitat de càlcul disposada.

**Title:** Traffic Matrices Space-Temporal Multiresolution Analysis

**Author:** Isaac Balasch

**Director:** David Rincón

**Date:** June, 15th 2010

## Overview

In this document it is shown the importance of traffic matrices, their characteristics and the ways they are treated at a mathematical level throughout diffusion models.

Traffic matrices, which are the volumes of traffic exchanged between any two nodes in a network, are very relevant for network engineering, specially regarding the research of predictive models.

The studies carried out with traffic matrixes, together with mathematic models such as diffusion wavelets, open the way for characterizing the behavior of a network, predicting it, and, what is of vital importance for network operators, to configure them in a proper way.

In the document some relevant mathematic concepts will be described, in order to justify the use of diffusion wavelets. Besides, the results of previous studies will be described and compared with the new techniques proposed in this work.

A diffusion operator based on the gravity model, and incorporating in the calculation a space-time correlation value, has been proposed.

It has been attempted to obtain an operator with better results than the ones achieved until now with other operators. The results obtained with data of Abilene (North American) and Geant (European) networks have been described from the mean squared error (MSE) perspective and from the compressibility of the energy of matrix in the fewest possible amount of coefficients.

Even though good results have been obtained they should be verified by further studies due to the lack of available computation capability.

# ÍNDIX

<b>INTRODUCCIÓ .....</b>	<b>7</b>
<b>CAPÍTOL 1. Matrius de Teletrànsit .....</b>	<b>8</b>
1.1. Definició.....	8
1.2. Importància de les Matrius de Teletrànsit.....	9
1.3. Aplicacions de les Matrius de Teletrànsit .....	9
1.4. Obtenció de les Matrius de Teletrànsit .....	11
1.5. Inferència de Matrius.....	12
1.6. Model de Gravetat.....	12
<b>CAPÍTOL 2. Model per Matrius de Teletrànsit basat en Anàlisi Multiresolució.....</b>	<b>15</b>
2.1. Model Dispers .....	15
2.2. Wavelets i l'anaàlisi Multiresolució (MRA).....	15
2.3. Wavelets per imatges 2D .....	18
2.4. Wavelets de difusió .....	19
2.5. Treballs previs: Wavelet de difusió en 2D .....	23
<b>CAPÍTOL 3. Dades Analitzades .....</b>	<b>29</b>
3.1. Descripció de les dades.....	29
3.2. Utilització de les dades .....	30
<b>CAPÍTOL 4. Correlació Temporal .....</b>	<b>33</b>
4.1. Coeficient de correlació.....	33
4.2. Anàlisi espai-temporal .....	38
<b>CAPÍTOL 5. Resultats.....</b>	<b>43</b>
5.1. Abilene.....	43
5.2. Geant.....	50
5.3. Comparativa Abilene vs Geant.....	52
<b>CONCLUSIONS.....</b>	<b>57</b>

**BIBLIOGRAFIA ..... 59**

**GLOSSARI..... 62**

## INTRODUCCIÓ

En aquest treball es realitzarà l'estudi de l'aplicació d'un operador de difusió sobre matrius de teletrànsit amb la intenció d'obtenir millors resultats que els obtinguts per altres operadors utilitzats fins el moment, respecte a la compressibilitat de la matriu en la mínima quantitat de coeficients (aquells que concentren la major part de l'energia de la matriu original).

En un primer terme es realitzarà una introducció a les matrius de teletrànsit, aclarint què són i la seva rellevància a la enginyeria de xarxes. A més, en el primer capítol es mostraran els diferents mètodes d'obtenció de dades per generar les matrius i les eines matemàtiques per analitzar-les.

En el segon capítol es mostrarà el tipus d'operacions aplicades sobre les matrius, realitzant una introducció al concepte d'anàlisi multiresolució. S'introduiran els models dispersos i la utilització de les wavelets per l'anàlisi multiresolució. S'explicarà l'evolució dels models matemàtics i la utilització de les wavelets per utilitzar-les en més d'una dimensió i a la seva vegada l'evolució per utilitzar-les en les matrius de teletrànsit amb les anomenades wavelets de difusió.

En el tercer capítol es mostrarà com és el format de les dades amb les quals s'ha treballat, les dues xarxes reals de les quals s'han obtingut les dades per fer l'anàlisi, i les seves característiques.

En el quart capítol es farà la descripció del operador de difusió utilitzat per realitzar l'estudi, amb el càlcul del coeficient de correlació utilitzat per introduir-lo a l'operador que analitzarà les matrius.

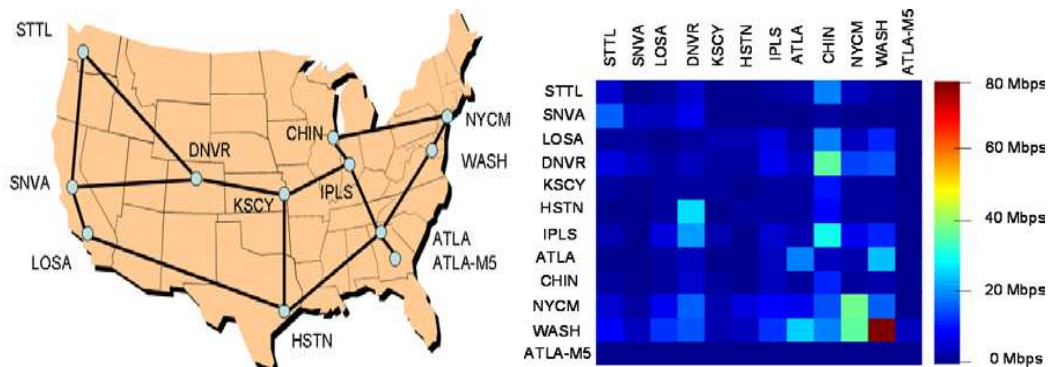
Al cinquè capítol es podran veure els resultats de l'estudi, on es podrà apreciar que depenent de les xarxes, els resultats són més o menys favorables. Això com s'explicarà, és possiblement degut a les diferents característiques de les dues xarxes.

Per finalitzar es presentaran les conclusions extretes de l'estudi, així com les línies futures que es poden seguir.

# CAPÍTOL 1. MATRIUS DE TELETRÀNSIT

## 1.1. Definició

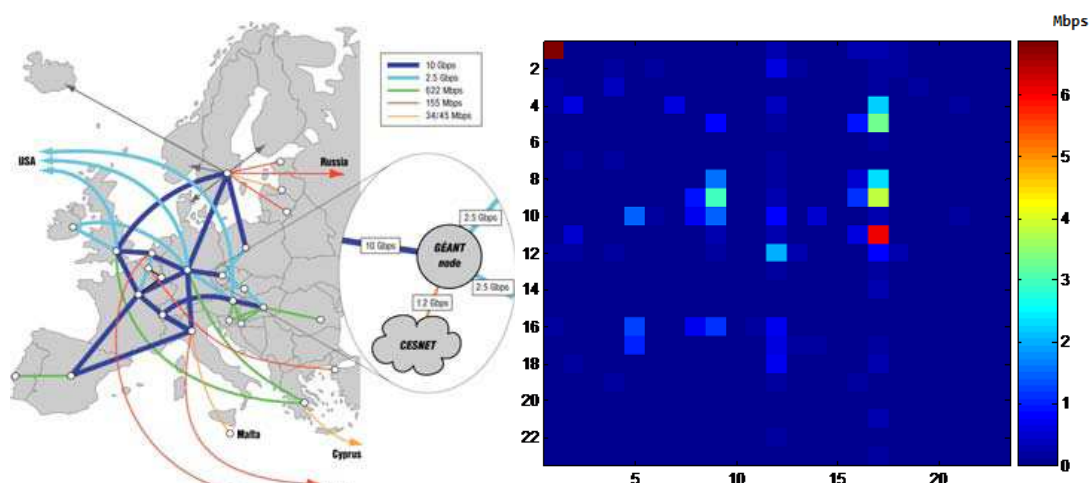
Les matrius de teletrànsit, (anomenades *traffic matrices, TM*) són una forma de descriure els volums de teletrànsit d'entrada i sortida intercanviats per nodes en una xarxa. Aquests nodes poden ser simplement routers o Points of Presence (PoPs), dels quals pegen altres nodes o xarxes senceres, (tipus LAN o MAN, per exemple). Per cada parell de nodes (entrada-sortida), les matrius de teletrànsit especifiquen la quantitat de teletrànsit que flueix sobre una ruta entre els nodes duran un determinat temps. En els nostres *datasets*, aquest interval és, per Abilene (la xarxa acadèmica nord-Americana), de cinc minuts i per a GÉANT (la xarxa acadèmica Europea), de quinze minuts. Les matrius no han de ser necessàriament simètriques, ja que la distribució del trànsit en una xarxa tampoc ho és. Un punt d'interès serà la generació de teletrànsit d'un node cap a ell mateix; això és degut a que en les matrius de teletrànsit tenim relacions per un origen X amb un destí X, és a dir parell X-X, (per exemple, Seattle-Seattle or LA-LA). Aquest fet és acceptable si el node és un PoP, ja que part del teletrànsit generat pels usuaris relacionats a aquest PoP anirà dirigit a altres usuaris que també estan connectats al mateix PoP



**Fig. 1.1.** Xarxa Abilene(USA) i representació gràfica d'una de les seves matrius de teletrànsit (3 de Març del 2004 des de les 12:00 a les 12:05).

A les figures 1.1 i 1.2 es pot apreciar el resultat de representar una matriu de teletrànsit com si fos una imatge, amb l'origen a l'eix vertical i la destinació a l'eix horitzontal.





**Fig. 1.2.** Xarxa GEANT (Europa) i una representació gràfica d'una de les seves matrius de teletrànsit (22 de Febrer del 2005 des de les 19:45 a les 20:00). No apareixen nodes perquè les dades estan anonimitzades.

## 1.2. Importància de les Matrius de Teletrànsit

Les matrius de teletrànsit tenen una gran utilitat potencial per la gestió de les xarxes IP i la capacitat de planificació a partir de la relació de tres paràmetres bàsics de la xarxa: origen, destí i volum de teletrànsit. Aquest tipus de matrius són utilitzades en enginyeria de teletrànsit a partir de les quals es pot configurar l'encaminament en termes de, per exemple, balanceig de carrega, qualitat de servei, etc.

Donada la importància de les TMs, la disponibilitat de bons models és essencial pels operadors de xarxes, ja que se'ls donarà bones estimacions matrius amb la capacitat d'aplicar aquestes estimacions a les seves operacions diàries.

Amb la informació que proporcionen es poden dissenyar noves xarxes, detectar anomalies, simulació de dades i fer prediccions de problemes a les xarxes en un temps determinat. Aquest és el motiu d'interès en els grups de recerca sobre la modelització de les matrius de teletrànsit i les seves prediccions.

## 1.3. Aplicacions de les Matrius de Teletrànsit

Els operadors han buscat sempre (i segueixen fent-ho), obtenir els millors models d'estimacions o solucions noves i més precises, als problemes que el flux de teletrànsit sobre un enllaç pot produir: planificació, fallades, obtenció d'informació per la monitorització, detecció *a priori* de fluctuacions impredecibles. Aquests són els principals objectius en l'anàlisi de TM's.

### 1.3.1. Síntesis

Un dels problemes que existeixen és la necessitat de grans quantitats de mostres de TMs per realitzar simulacions fiables, per exemple, d'una nova xarxa o un nou protocol. Una vegada obtinguda una síntesi de TM, es poden crear matrius de teletrànsit amb les quals es dissenyen i creen xarxes senceres (localització de routers o PoPs, enllaços i assignació de capacitats). La síntesis ha de ser tant fiable com sigui possible, amb l'objectiu de minimitzar l'error entre la síntesis i les mostres originals [17,18].

### 1.3.2. Routing i balanceig de càrrega

Una altra aplicació important de les TM's és l'optimització del encaminament. A la enginyeria de teletrànsit es busca l'optimització del rendiment del encaminament en termes de balanceig de càrrega, qualitat de servei (QoS), etc. Podrien existir enllaços amb un gran càrrega agregada que suposés una congestió de la xarxa: això podria ser solucionat introduint rutes alternatives per balanceja la càrrega cap a altres nodes de la xarxa. Per aquest motiu, seria important tenir algun model per fer prediccions sobre les variacions de la demanda a un llarg i mitja termini. Amb aquesta informació l'operador podria reconfigurar la xarxa per controlar pics de teletrànsit.

### 1.3.3. Detecció d'anomalies

Existeixen diferents anomalies en les que un operador de xarxa pot està interessat en detectar. Per exemple, la probabilitat que un grup d'usuaris (o només un) mostri un canvi de comportament, per exemple, incrementant la seva demanda de teletrànsit. En aquest cas, el model hauria de ser capaç de senyalar que alguna cosa està succeint, per a que la xarxa pugui donar resposta. Una altra possible anomalia seria el fet que un usuari hostil o un error de configuració a la xarxa realitzes una gran quantitat de peticions a un node. Aquest fenomen és l'anomenat atac de denegació de servei, *-Denial of Service attack (DoS)-*, i en el cas que fos provocat intencionadament, és utilitzat per bloquejar algun servei o host per sobrecarrega de teletrànsit d'entrada. En aquest cas el model hauria de mostra que algun fet esta sortint de la pauta habitual. Respecte al balanceig de càrrega, la detecció d'aquestes anomalies poden ajudar a reconfigurar les rutes de la xarxa i las seves càrregues.

### 1.3.4. Predicció d'anomalies

Un dels àmbits de recerca més importants és trobar un model que proporcioni una predicció fiable de les futures TMs, per poder comparar-les amb les reals i així poder detectar les anomalies. El fet d'avançar-se els canvis, i actuar abans que aquest succeeixin serà un gran assoliment. Desafortunadament, no existeixen grans coneixements sobre com obtenir un model fiable, això suposa que molts treballs de recerca estiguin centrats en aquest àmbit [1,6].

## 1.4. Obtenció de les Matrius de Teletrànsit

Com s'ha vist, les matrius de teletrànsit són una mesura de gran importància per la enginyeria de teletrànsit i la panificació de xarxes. Tot i la importància esmentada apareixen problemes per obtenir-les. El primer problema és com realitzar mesures de dades en temps real. Per realitzar això es disposant de dues tècniques, Netflow i els comptadors SNMP.

### 1.4.1. Netflow

Netflow [28] és una aplicació de xarxa desenvolupada per Cisco, però existeixen productes similars d'altres fabricants. Entre d'altres característiques aquest protocol pot mesura fluxos i volums de teletrànsit que són commutatats per un router. El principal problema és que quan un node rep gran quantitat de fluxos requereix gran part de la capacitat de la CPU per processar la informació rebuda, en detriment de les tasques d'encaminament, pel que és complicat el seu ús en MANs o WANs (Xarxes d'àrea Metropolitana o Esteses), com les que s'han utilitzat per aquest estudi. Un altre raó és la dificultat per realitzar mesures en temps reals en xarxes que poden tenir centenars de nodes, ja que suposaria un compromís pel consum de les CPUs

### 1.4.2. SNMP (*Simple Network Management Protocol*)

SNMP és un protocol de xarxa utilitzat pel monitoratge d'equips. La millor característica del SNMP és que està present a totes les targetes de xarxa, el que permet el fet de no tenir determinats elements per las mesures. A més aquest protocol no afecta en gran mesura las tasques de routing com el Netflow, degut a que treballa amb MIBs (*Management Information Bases*), les quals requereixen un cost computacional molt baix. En aquestes bases de dades d'emmagatzematge d'informació es poden trobar gran quantitat de dades transmises per un enllaç, sense tenir en compte la destinació final, amb una regularitat típica de cinc minuts. De totes maneres, el problema és que les dades que s'obtenen son el valor agregat de tots els fluxos que passen a través de l'enllaç, pel que és necessari realitzar un procediment d'inferència per diferenciar els diferents fluxos.

El segon problema prové del fet que els ISP comercials no volen publicar aquestes dades. Aquest fet és comprensible degut a la competència entre empreses, per un altre banda aquest és un dels motius principals en la lentitud en la investigació en aquest camp. Els investigadors en algunes ocasions se'ls permet la captura de mostres de temps reals per les pròpies organitzacions. De totes maneres conjunts de dades de llarga duració són molt difícils d'obtenir. S'ha d'agrair a Zhang i Uhlig [1, 15], per la publicació d'un parell de conjunts de dades de GEANT i Abilene que són els utilitzats en els nostres experiments, i són el punt de partida per la investigació de models.

## 1.5. Inferència de Matrius

Amb SNMP s'obté informació sobre el teletrànsit per cada un dels enllaços, però a diferència amb el Netflow, amb el que s'obtenen dades de cada flux, amb SNMP no s'obté tota la informació, i en conseqüència per completar les matrius s'ha de realitzar algun tipus d'inferència. Aquesta falta de dades reals es pot solucionar fent una estimació de les dades no disponibles. Per aquesta estimació s'ha utilitzat la següent expressió:

$$\mathbf{y} = \mathbf{A} \mathbf{x} \quad (1.1)$$

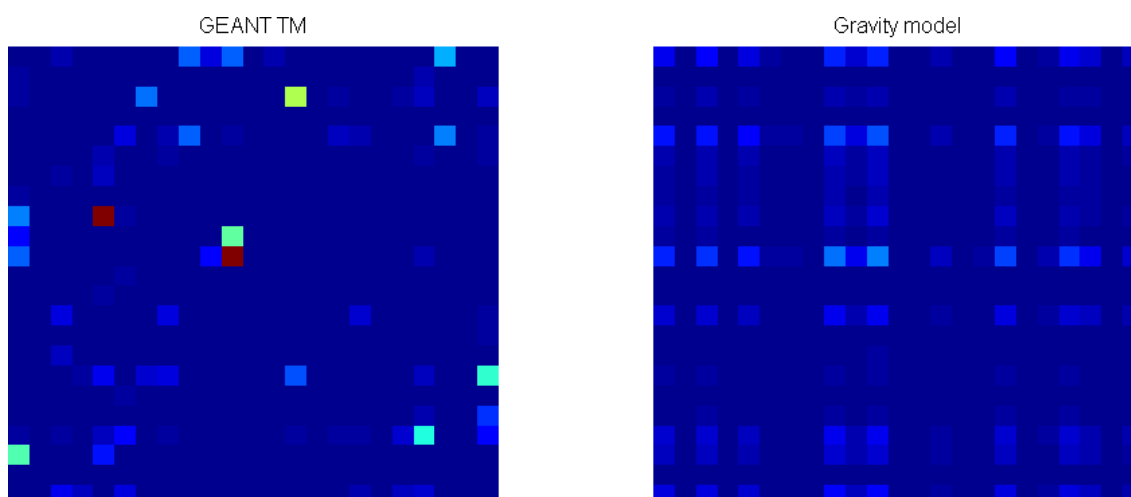
on  $\mathbf{y}$  representa el vector de comptadors SNMP d'enllaços (de longitud  $K$ , on  $K$  representa la quantitat d'enllaços a la xarxa),  $\mathbf{A}$  és la matriu de routing ( $a_{ij}=1$ , si l'enllaç  $i$  pertany a la ruta associada al parell  $j$  origen - destí), amb dimensions  $K \times N^2$ , sent  $N$  el número de nodes, i  $\mathbf{x}$  és la matriu de teletrànsit (escrita com un vector  $N^2 \times 1$ ,  $x_{ij}$  és el valor del teletrànsit associat amb el parell  $i-j$  origen - destí). Desafortunadament com  $K$  és un valor molt més petit que  $N^2$  deduir una matriu de teletrànsit a partir de les mesures SNMP es converteix en un problema invers irresoluble amb infinites solucions i per tant es necessiten dades addicionals per obtenir una solució. Un dels models que aporten informació addicional és el Gravity Model (Model de Gravetat).

## 1.6. Model de Gravetat

El Model de Gravetat està basat en la llei de la gravetat de Newton, la qual indica que la força d'atracció entre dos cossos és directament proporcional al producte de les seves masses i una constant gravitacional, i inversament proporcional al quadrat de la distància que hi ha entre els dos cossos. Aquesta idea també ha estat utilitzada a les ciències socials i al transport, on trobem models de gravetat per les transaccions econòmiques o també en fluxos migratoris. En tots aquests casos, la formulació del model pot portar a una bona analogia entre el paràmetres originals per massa i distància i uns paràmetres adaptats per cada interpretació de la llei de la gravetat. El model de gravetat depèn del producte de les masses (població, terminals) de dos cossos (ciutats, nodes), una valor constant (constant gravitacional universal en la formulació original) i la distància (correlació, adjacència). En el cas de migracions de població la tendència és de migrar a grans ciutats (la més gran de les masses, la força d'atracció més forta; en el cas de la economia, el mercats econòmics gegants concentrant els negocis).

En el nostre cas, el model de gravetat assumeix la hipòtesis que el teletrànsit que circula entre dos ciutats amb gran població serà més dens que el que circula entre dos ciutats amb menys població. Com la força gravitacional, la quantitat de dades que circula entre Madrid i Barcelona serà elevada, degut a que existeix més "població" (terminals) les quals utilitzant els serveis d'un ISP. Per contra les dades que circulen entre Castelldefels i Barcelona haurien de ser menys intenses que les primeres entre Barcelona i Madrid.

Per calcular el model de gravetat per una TM, no es considerarà la distància física (el paràmetre original de la llei de la gravetat referit a la distància no és tan evident en aquest model, els algoritmes de “routing” a les xarxes poden encaminar el paquets a través de camins alternatius, per tant la distància no pot fer referència a la distància física), i s'utilitzaran volums de teletrànsit d'ingrès-sortida en comptes d'utilitzar directament la població. Amb això s'obtidran 2N paràmetres (N per la fracció de teletrànsit total generat per cadascun dels nodes i N més per la fracció del teletrànsit rebut total per cada node).

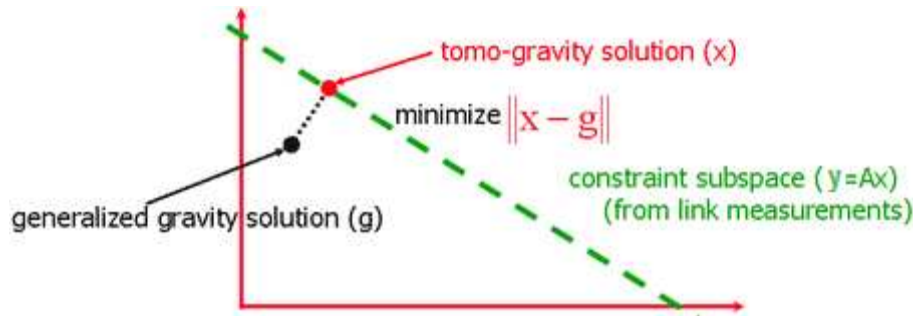


**Fig. 1.3.** Model de Gravetat (GM). Esquerra: TM de GÉANT (22 de Febrer del 2005 , a les 13:00). Dreta: Model de Gravetat de la matriu prèvia.

Quan s'obté la matriu de gravetat, es conserva la totalitat del teletrànsit : es pot afirmar que la energia es conserva. Matemàticament  $GM = T_i \times T_o^T \times \text{Volum Total}$  (on  $T_i$  és el vector de probabilitats d'entrada i  $T_o$  és el vector probabilitats de sortida del teletrànsit per a cada node).

En aquest punt es pot introduir el concepte de tomogravetat per visualitzar com el model d'inferència és utilitzat en el escenari d'inferència de TM. D'això se'n dedueixen dos qüestions, la primera, tenint el model de Gravetat com una solució inicial s'utilitzarà cada dada carregada en els “edge link “ i la política de “routing” del ISP, llavors es farà una estimació tomogràfica, trobant una redefinició de la solució inicial aplicant alguna mètrica de distància per trobar la solució (en el subespai de solucions de l'equació  $y = Ax$ ) més propera al model original. Es pot veure un exemple a la Fig. 1.4.

Per finalitzar, fer notar que a [13], els autors informen que el mètode és bo i proporciona un  $\pm 20\%$  d'error obligat (mesurat com error quadràtic mig, MSE, entre la TM original i la mateixa trobada amb el procediment de tomogravetat).



**Fig. 1.4.** Solució de mínims quadràtics. Podem apreciar la projecció des del model de gravetat al subespai de solucions de l'equació 1.1 [23].

## CAPÍTOL 2. MODEL PER Matrius DE TELETRÀNSIT BASAT EN ANÀLISI MULTIRESOLUCIÓ

### 2.1. Model Dispers

Un dels objectius a llarg termini dels nostres estudis és desenvolupar models per les TMs. L'objectiu és que els models siguin dispersos, en el sentit de tenir menys paràmetres que els de la dimensió real de la pròpia matriu. En una xarxa amb  $N$  nodes la TM tindrà  $N^2$  elements. En xarxes grans  $N$  pot ser de l'ordre de centenars o milers de nodes, mentre que el model dispers només tindrà  $M$  elements que seran molt inferiors a  $N$ .

Un model matemàtic ha de satisfer una sèrie de característiques [6]. Per exemple, hi ha un compromís entre la capacitat predictiva del model i la seva fidelitat: si es tenen una gran quantitat de coeficients d'un conjunt, s'obtenen models més específics, però aquest model no servirà per altres grups de dades. Per una altra banda, si es construeix un model amb poques dades de diferents conjunts, s'aconseguirà un model predictiu per tots aquests conjunts, però no serà tant ajustat a cap dels conjunts, per tant els errors de les prediccions poden ser elevats.

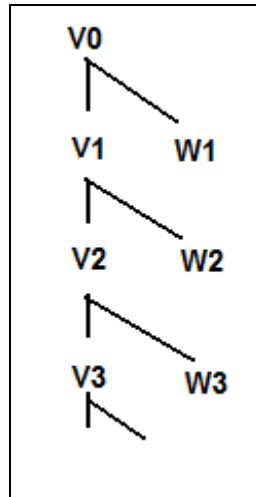
Si el model té pocs paràmetres la compressibilitat de les dades serà elevada, per tant es podran estudiar certes característiques amb poca capacitat de càlcul. D'altra banda, això redundarà en una alta compressibilitat de les dades de la matriu, que podran ser representades amb pocs coeficients que concentrin la seva energia.

Per finalitzar, es podrien obtenir millors solucions per el problema d'inferència mencionat al capítol anterior, on tenim  $K$  mesures (on  $K$  de l'ordre de  $N$ ) i  $N^2$  paràmetres per deduir (per tant més incògnites que equacions). Aquest problema es podria solucionar si el model tingues  $M$  valors que siguin molt inferior al número de valors  $K$ .

### 2.2. Wavelets i l'anaàlisi Multiresolució (MRA)

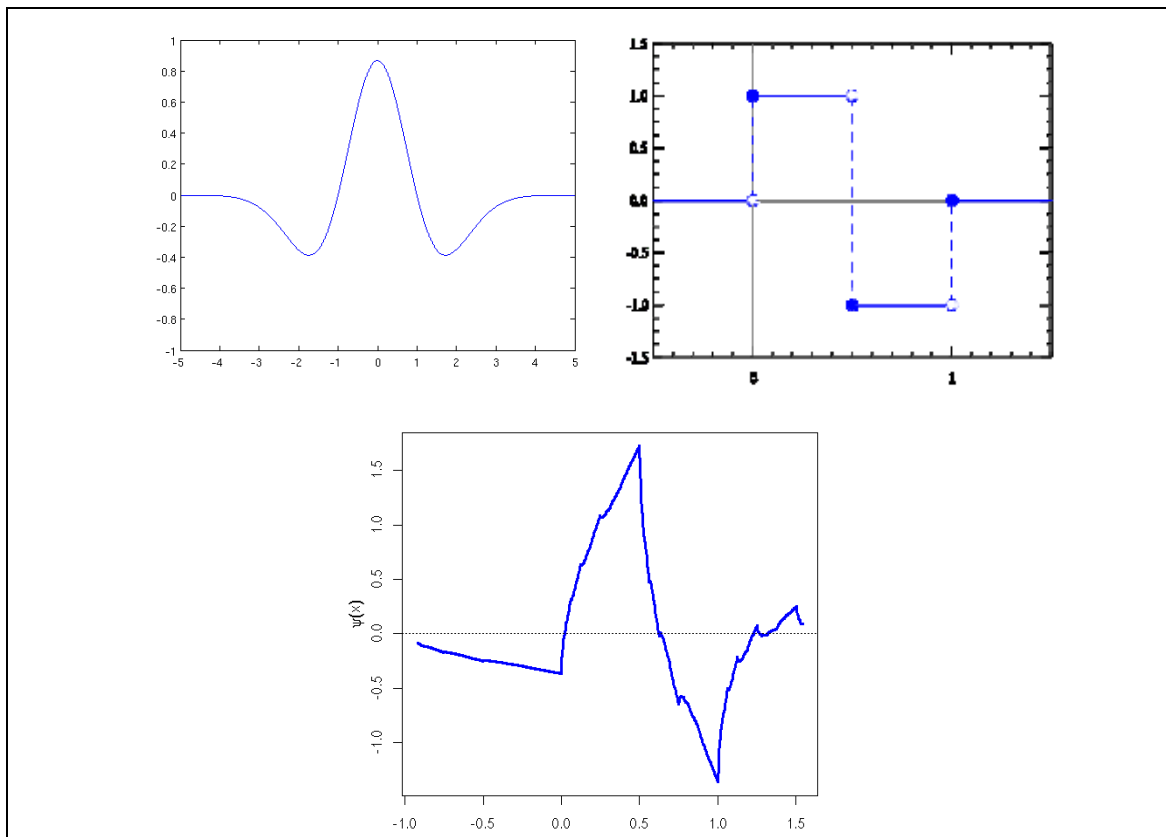
La tècnica utilitzada per obtenir un model dispers és la de l'anàlisi multiresolució. L'anàlisi Multiresolució (MRA) permet analitzar un conjunt de dades *-dataset-* a diferents escales de detall.

Matemàticament, MRA [29] consisteix en obtenir d'un grup de subespais anidats  $V_j$  (aproximacions) i els seus complements ortogonals, un altre grup de subespais anidats  $W_j$  (detalls), el quals verifiquen que la diferència entre els dos subespais d'aproximacions consecutives són el detall de l'escala  $j$  ( $W_j = V_{j-1} - V_j$ ). Per les aproximacions  $V_j$  (per  $j \geq 1$ ) sempre hi haurà la possibilitat per descompondre a una determinada escala els subespais d'aproximacions i els detalls. De forma inversa es pot reconstruir un espai escalat si es tenen els subespais d'aproximacions i els dels detalls.



**Fig. 2.1.** Exemple d'anàlisi MRA. En el moment de començar, les senyal original és dividida en detalls ( $W$ ) i aproximacions ( $V$ ). En el següent pas, aquestes aproximacions seran descompostes denou en detalls i aproximacions.

La transformada Wavelet [16,20] és una de les tècniques preferides per desenvolupar un MRA. Es pot entendre una wavelet com una funció matemàtica utilitzada per dividir la senyal en diversos components d'escala, per exemple, la senyal es pot representar en versions traslladada i dilatada d'una forma d'ona bàsica anomenada "wavelet mare".

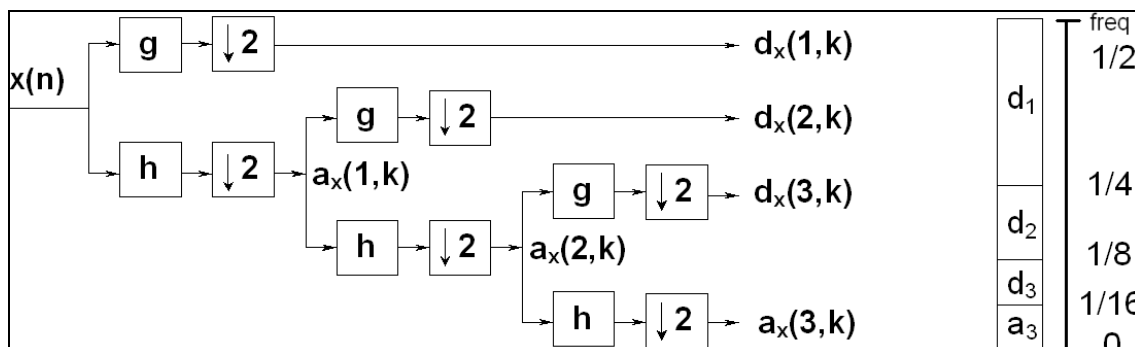


**Fig. 2.2.** Exemple de wavelets mare: Esquerra "Mexican hat", Dreta "Haar", Abaix "Daubechies".



En el camp de processament de senyal és habitual utilitzar mètodes basats en wavelet per comprimir i eliminar soroll de les senyals, series temporals o imatges [16]. La transformada Discreta Wavelet (DWT) [25] analitza els senyals a través del seu producte escalar amb dues funcions base anomenades wavelet pare (o funció d'escalat  $\varphi(t)$ ) i wavelet mare ( $\psi(t)$ ), amb longitud finita. Aquestes són dilatades, en potències de 2, y traslladades per cobrir la totalitat del domini de les senyals originals, obtenint l'anàlisi de la senyal pels instants  $t = 2j - k$  (on  $j$  és la escala i  $k$  és el desplaçament temporal).

L'objectiu de la funció d'escalat és capturar freqüències baixes del senyal, mentre que les altes freqüències (o detalls) són analitzades per la wavelet mare. Si la funció base compleix certes condicions, la transformada resultant és ortonormal i pot ser implementada amb un filtre pas-baix i un filtre pas-alt ( $h(n)$  i  $g(n)$ , respectivament, relacionats amb  $\varphi(t)$  i  $\psi(t)$ ), com es mostra a la figura 2.3. Els filtres pas-baix mostren les successives aproximacions de les escales a mesura que van sent menys detallades. Es poden interpretar com una imatge que va passant pels filtres pas-baixos. A cada iteració els filtres pas-alt capturen els detalls de les altes freqüències:  $d_x(j,k)$  agafa les diferències entre  $a_x(j-1,k)$  i  $a_x(j,k)$ , on  $j$  és un nivell més baix (o més borrós), de descomposició que  $j-1$ . Aquest mètode permet recuperar la senyal original únicament repetint el procés de forma inversa: sintetitzant aproximacions i detalls a partir dels coeficients de la transformada wavelet.



**Fig. 2.3** : Esquerra: filtres pas alt i pas baix en cadena per a una DWT i  $j=3$  escales, obtenint les aproximacions  $V_3 a(3,K)$  i els detalls  $W_3 d(3,k)$  per a  $j = 1,2,3$ . Dreta: descomposició del espectre subseqüent normalitzat.

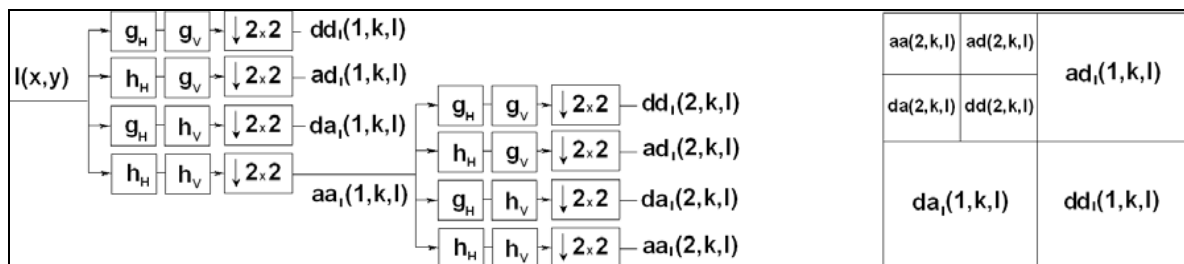
En el domini freqüencial, DWT crea una descomposició per subbandes, on l'espectre es tractat a cada escala. Això produeix un anàlisi multiresolució on el senyal original és descompost a la seva aproximació final (per la freqüència més baixa que es mantindrà per la mitja de la senyal) i un grup de detalls a diferents altes freqüències (coeficients wavelet).

En resum, la senyal  $x(n)$  és descomposta inicialment amb un parell de filtres. El senyal obtingut a la sortida del filtre pas-alt ( $g$ ), és conegut com els detalls  $d1$ . El primer senyal obtingut a la sortida del filtre pas-baix ( $h$ ) és l'aproximació  $a1$ . A la segona passa, es pot descompondre l'aproximació  $a1$  en els seus detalls  $d2$ , i aproximacions  $a2$ . El senyal  $a2$  pot ser descompost de nou en les seves aproximacions i detalls.

Aquest procés pot ser repetit tantes vegades com es requereixi. Una vegada descomposta la senyal es pot realitzar el procés invers per recuperar la senyal original. Si es descarten alguns coeficients dels detalls s'obté una senyal aproximada.

### 2.3. Wavelets per imatges 2D

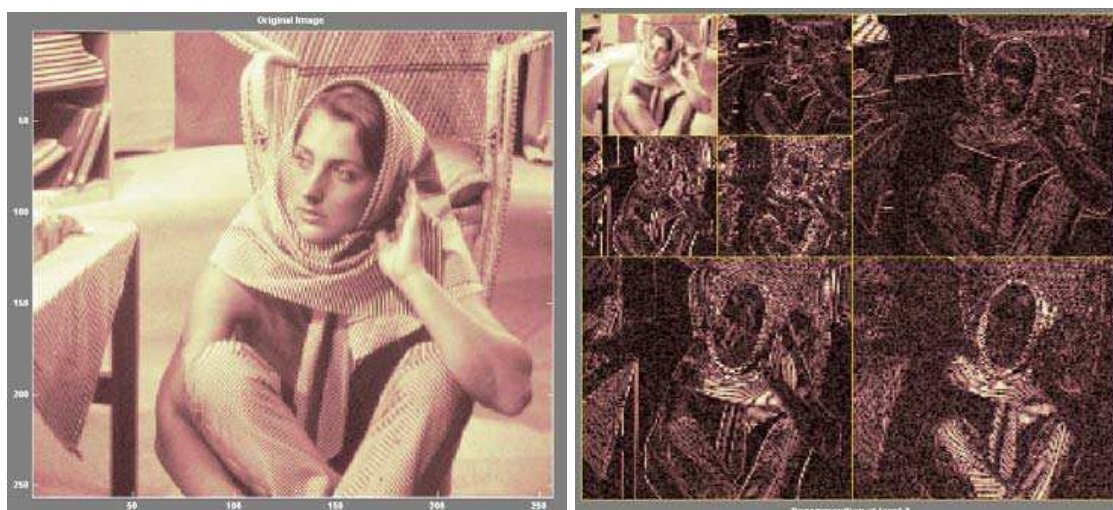
Un dels camps on les wavelets han tingut una contribució important és en el processament i compressió d'imatges. Per exemple, JPEG 2000 [30] utilitza les wavelets 2D per reduir la quantitat de bytes necessaris per codificar una imatge, mantenint una qualitat acceptable.



**Fig. 2.4.** : Esquerra, filtres pas alt i pas baix en cadena usats en el procés wavelet en 2D. A la dreta la descomposició del espectre normalitzat. Veure la diferència amb la wavelet d'1D a la figura 2.3.

Les Wavelets 2D són una generalització de les wavelets clàssiques calculades en els eixos  $x$  i  $y$  d'unes dades bidimensionals. La forma en la que treballa és similar al procés explicat prèviament per senyals d'una dimensió. L'únic que s'ha de tenir en compte és la fragmentació de cada subespai en quatre, cada un correspon a les quatre combinacions dels filtres pas-alt i pas-baix,  $(h(n)-h(n))$ ;  $h(n)-g(n)$ ;  $g(n)-h(n)$ ;  $g(n)-g(n)$ , en els dos eixos verticals de la imatge, per tant, es tenen quatre combinacions de filtres pas-alts i pas-baix, com es pot apreciar a la figura 2.4.

Com es pot veure a l'exemple de la imatge real de la figura 2.5, a la imatge descomposta es pot apreciar la imatge aproximada al requadre petit a dalt a l'esquerra, a la resta de requadres es pot apreciar els detalls de la primera escala (requadres més grans) i els de la segona escala. Per a cada cas s'obtenen tres sortides, las quals estan formades per la combinació dels filtres alt-baix, baix-alt i alt-alt. En aquestes imatges es pot notar que els detalls generen la construcció de la silueta de la imatge. Si s'aplica JPEG 2000, es podrien descartar alguns dels coeficients del detalls i mantenir una aproximació acceptant una certa pèrdua de qualitat.



**Fig. 2.5.** : Esquerra: exemple d'imatge original; Dreta: descomposició wavelet en 2D de la imatge. Exemple de Matlab.

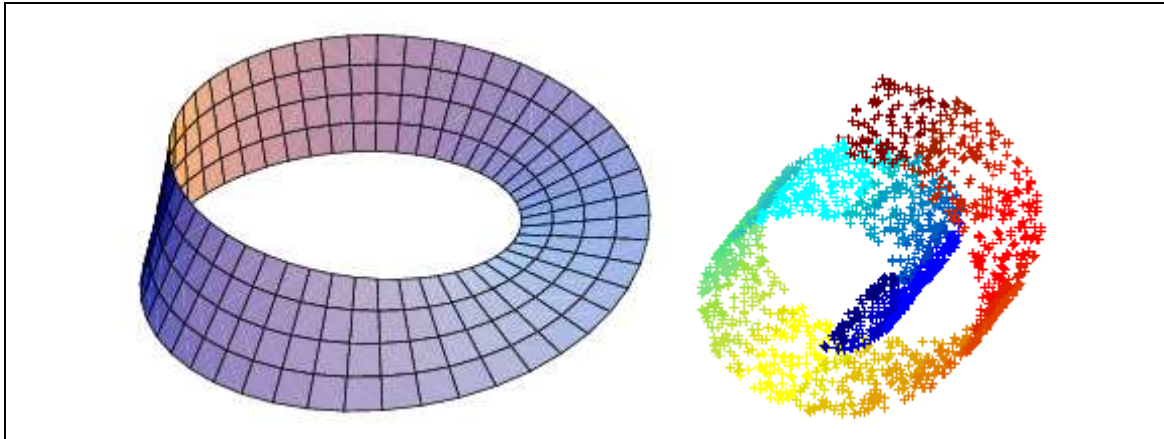
## 2.4. Wavelets de difusió

Si volem fer MRA sobre xarxes, hem d'estendre l'anàlisi multiresolució a grafs. El problema està en que no es pot aplicar el procés utilitzat amb les imatges per matrius de teletrànsit. La relació entre els dos nodes d'un graf, (una xarxa amb terminals o nodes i enllaços entre ells), no és exactament el mateix que la relació entre píxels d'una imatge, ja que un graf és molt més complex.

En las imatges (per exemple JPEG), part de la compressió ve donada per inferència: el valor d'un píxel pot ser obtingut per el valor dels píxels propers. Si un píxel és negre, es pot assumir amb una gran probabilitat que el píxel que està situat a la seva dreta serà negre, o com a mínim obscur. Aquesta relació no ocorre a las matrius de teletrànsit. Las matrius de teletrànsit es defineixen sobre un graf, i els graf no satisfan les condicions dels píxels, encara que es visualitzin las matrius com una imatge. Dos píxels de la imatge són adjacents però a la xarxa poden representar dos nodes que estan a dos extrems diferents de la xarxa. Per tant es necessiten eines diferents que permetin aplicar l'anàlisi multiresolució sobre un graf. Per a realitzar aquest anàlisi es necessiten diferents classes de wavelets

Una primera aproximació, desenvolupada per Crovella i Kolaczyk [5], introdueix les *Graph Wavelets* com una extensió per la transformada wavelet 2D. Això permet calcular la diferència entre las càrregues dels enllaços en nodes separats per un cert nombre de salts: el concepte d'escala és reemplaçat per la distància de salts entre enllaços. Els autors també mostren com els grafs wavelet poden ser usats per la detecció d'anomalies. El principal problema és la manca d'un algoritme computacional ràpid, i que la transformada no és ortonormal. A més els grafs wavelet no mostren el teletrànsit de manera dispersa: s'obté una descomposició redundat que és similar al resultat obtingut amb la transformada wavelet continua.

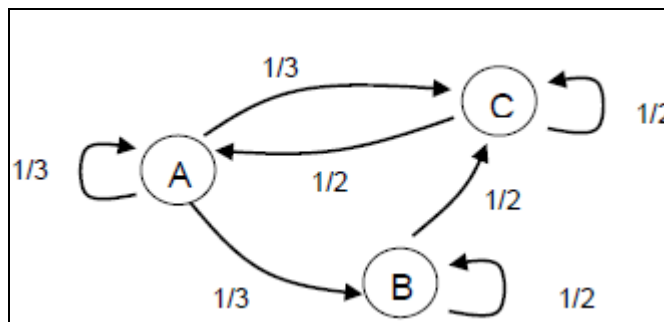
En aquest punt una altra generalització de wavelets entra en joc, la transformada Wavelet de Difusió (DW) [6,9] permet desenvolupar l'anàlisi multiresolució sobre un graf o un "col·lector" (*manifold*). Intuitivament, un "col·lector" és un espai matemàtic on dos punts del mateix cos poden estar molt propers a l'espai Euclidià, però la distància sobre la superfície del cos pot ser molt més gran.



**Fig. 2.6.** : Tira de Möbius (esquerra) i rotllo Suïss (dreta): exemples de "col·lectors". Dos punts són propers en l'espai però llunyans al llarg del cos.

Les Wavelet de Difusió, desenvolupades per Coifman i Maggioni [9], s'han convertit en una eina molt important, ja que permet realitzar un MRA sobre un graf. La primera passa és definir el graf i l'operador de difusió, on aquesta tindrà el mateix paper que la funció d'escalat a las wavelets clàssiques. Aquest operador està definit per una matriu i pot ser entès com la base sobre la qual la funció és projectada, o com la columna vertebral de la difusió.

La idea que hi ha darrera de l'operador de difusió és la d'explorar el graf (distàncies, adjacències ... ), per mitja de la difusió induïda per l'operador. En els primers experiments [10,34] es va utilitzar un operador "random walk" sense ponderació . El "random walk" té en compte els nodes adjacents i els hi dona un pes igual a l'inversa de la quantitat d'enllaços que existeixen en cada node. Si es veu el graf com una cadena de Markov, la probabilitat de "salt" a través de cada enllaç és la mateixa (equiprobabilitat). Amb això s'obté una matriu simètrica i estocàstica per l'operador. En termes de teoria de grafs, la probabilitat de sortir d'un node a través d'un cert enllaç és l'invers del grau del node. Aquest operador agafa l'encaminament dels fluxos o altres propietats sobre la distribució del teletrànsit, això és únicament un primer exemple. La següent passa és dilatar l'operador per calcular las seves potències: les  $n$ -èssimes potències del "random walk" en termes de la cadena de Markov associada són las probabilitats  $P_{ij}$  de la sortida d'un cert node  $i$  a un node  $j$  en  $n$  passes o salts. A mesura que es prenen potències més altes, tendint a infinit, el que s'obté és la probabilitat d'anar d'un node a un altre en qualsevol nombre de salts (és a dir, en estat estacionari), i la interpretació és que l'operador s'ha "difós" sobre el graf.



**Fig. 2.7:** Exemple de “randomwalk” sense ponderar en una xarxa amb tres nodes.

Assumint que una funció donada està definida pels vèrtex del graf, quan l'operador és aplicat a la funció, difumina la funció donada en el graf. El resultat és que els nodes més propers es van barrejant i confonen més ràpidament que els nodes més llunyans, (on la distància està definida en termes del graf).

Al treballar sobre un graf en comptes d'una imatge (desenvolupant l'MRA), l'analogia resultant de la descomposició DWT serà una selecció de subbandes en termes del operador de difusió dels autovalors i autovectors. És ben conegut [19,24] que un operador lineal T pot ser expressat com :

$$T = \sum_{i=1}^n \lambda_i v_i^T v_i \tag{2.1}$$

on  $\lambda_i$  són els autovalors de T i  $v_i$  són els autovectors associats. S'ha utilitzat una versió normalitzada de T, per que el valor més gran fos u. Unes altres propietats interessants són per exemple:  $\lambda^n$  són els autovalors per  $T^n$  com  $|\lambda| < 1$  el creixement de n ho farà  $|\lambda|^n$  tendint a zero. Els autovalors estan relacionats amb les freqüències incloses a cada subbanda. També podem veure un paràmetre addicional  $\epsilon$  que determinarà la precisió de la descomposició. La idea és classificar els autovalors propis comparant-los amb  $\epsilon$ , llavors els autovalors per sota d' $\epsilon$  seran identificats com a detalls, mentre que els que estiguin per sobre seran las aproximacions. Aquest procés serà repetit pas a pas fins que tots els autovalors menys un estiguin per sota del llindar marcat per  $\epsilon$  [10].

S'ha de recordar que els autovalors seran la solució de l'equació:

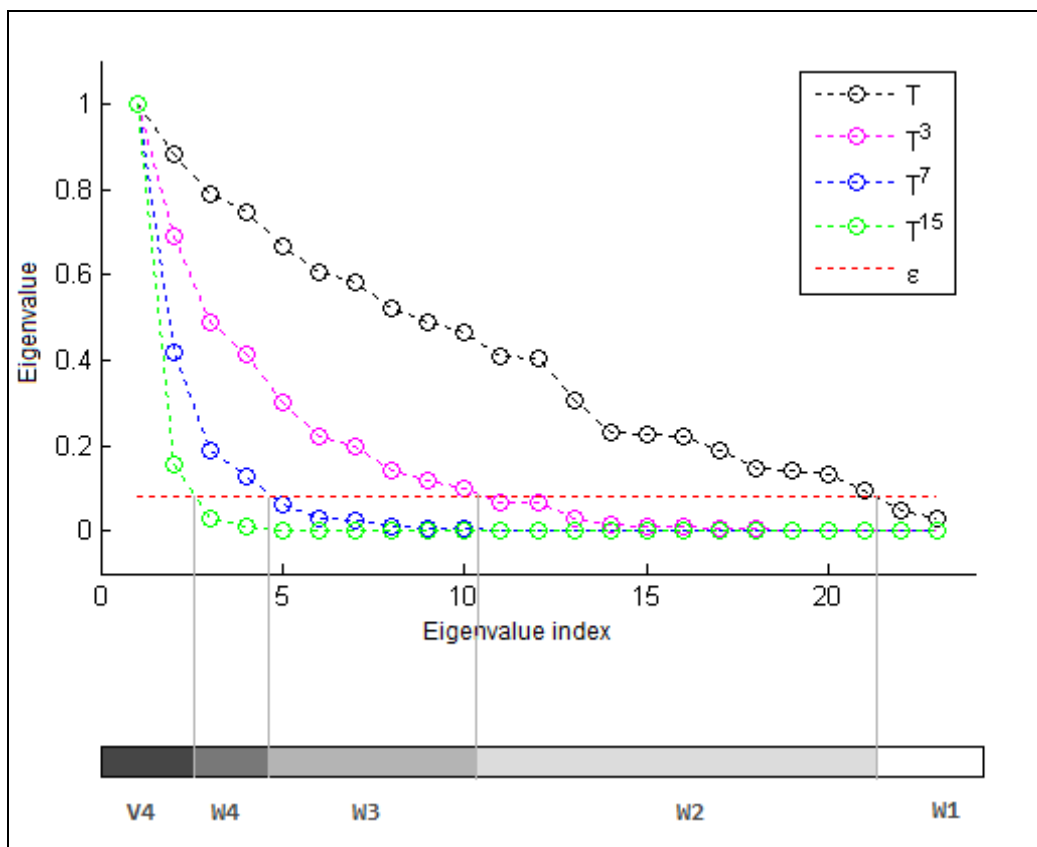
$$(A - \lambda I) x = 0 \tag{2.2}$$

on **A** és la matriu del operador de difusió, **I** és la matriu identitat,  $\lambda$  és un autovalor i **x** és l'autovector associat. Per a una matriu (de l'operador de difusió) n x n, es tindran n solucions (n autovalors) i es poden deduir algunes coses dels grafs des dels seus propis autovalors. Per exemple, el numero de autovalors zero són els numero de components connectats en el graf, i  $\lambda_1$  és l'anomenada connectivitat algebraica. Aquests autovalors son també útils per determinar famílies de grafs [6,19,24].

L'expressió 2.2 pot ser aplicada al teorema espectral [24] ja que sota condicions específiques una transformació lineal d'un vector  $\mathbf{v}$  pot ser expressada com una combinació dels seus autovectors ( $\mathbf{v}_i$ ) i els seus autovalors ( $\lambda_i$ ):

$$T(\mathbf{v}) = \lambda_1(\mathbf{v}_1 \cdot \mathbf{v})\mathbf{v}_1 + \lambda_2(\mathbf{v}_2 \cdot \mathbf{v})\mathbf{v}_2 + \dots + \lambda_{n-1}(\mathbf{v}_{n-1} \cdot \mathbf{v})\mathbf{v}_{n-1} \quad (2.3)$$

Després de tenir normalitzat l'operador per el seu valor més elevat, quant s'agafen les  $n^{\text{th}}$  potències de  $T$  ( $T^n$ ),  $n$  tendint a  $\infty$ , tots els autovalors seran zero, a excepció del més gran que serà u (aquest serà 1, ja que els autovalors estan normalitzats). En aquest moment només es tindrà un autovalor per sobre del llindar  $\varepsilon$ . D'aquesta forma, i aplicant l'operador de difusió una vegada i una altra, s'obté un espectre dividit del graf, llavors, els autovalors mantinguts expandiran els seus subespais de baixes freqüències (detalls). A cada pas, els autovalors són ortonormalitzats de nou.



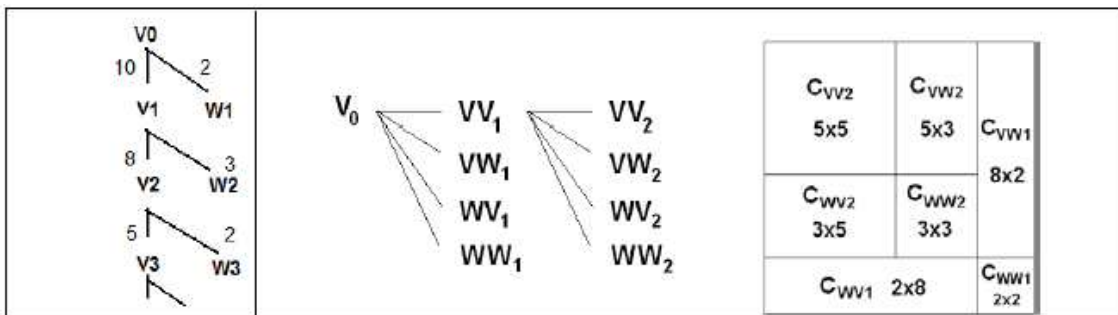
**Fig. 2.8.** Autoespectre de les potències d'un operador  $T$ . Es pot apreciar que a mesura que  $n$  augmenta més autovalors són descartats, per exemple assignant als detalls del subespai  $W$ . Quan  $n \rightarrow \infty$ , només el valor més elevat (1) es manté.

Una vegada s'han obtingut els subespais d'aproximacions i detalls ( $V_i$  i  $W_i$ ) es pot projectar qualsevol funció  $F(v)$  (definida sobre els vèrtex del graf) sobre els subespais i obtenir un MRA de  $F(v)$ . Aquest procés té un cost computacional molt elevat, però Coifman i Maggioni [9] van desenvolupar un algorisme ràpid per calcular-ho i obtenir els coeficients dels detalls i les aproximacions.



### 2.5. Treballs previs: Wavelet de difusió en 2D

El teletrànsit de les matrius es pot representar com una funció bidimensional  $F(x,y)=z$  on  $x$  i  $y$  són els nodes d'ingrés i sortida en aquest ordre i  $z$  representa el volum de teletrànsit. Tenint en compte que la transformada de Wavelet de Difusió només es pot aplicar a una funció d'una única dimensió  $F(x)$ , s'ha d'estendre la Wavelet de Difusió a 2D i continuar amb la mateixa aproximació de la DWT en 2D [6]. S'ha pogut veure anteriorment en aquest document que el procés per estendre la DWT a 2D: descomposició en quatre subbandes (un parell per a cada dimensió) per a cada escala i repetint el procés a la següent pels subespais de les aproximacions (resultat dels filtres pas baix - pas baix). En termes generals, aquest és el mateix procés, on ara s'obtiniran tres subbandes de detalls i només una aproximació. Anàlogament, la transformada DW en 2D seguirà projectant  $F(x,y)$  una vegada a cada "dimensió" (origen i destí) per obtenir un producte tensor de base 1D tant per les aproximacions com pels detalls. Els detalls de l'algoritme varien, però el procés és intuïtivament similar al de les wavelets clàssiques en 2D. En aquest cas els coeficients  $C_{VV1}$ ,  $C_{VW1}$ ,  $C_{WV1}$  i  $C_{WW1}$  generen les subbandes  $VV_1$ ,  $VW_1$ ,  $WV_1$  i  $WW_1$  (on  $VV$  és l'aproximació del subespai de freqüències baixes-baixes,  $WW$  és el subespai de detalls de les freqüències altes-altes, i  $VW$  i  $WV$  són les combinacions baxa-alta i alta-baixa). La transformada en 2D és també ortonormal i invertible, fet que permet reconstruir la funció original a partir dels coeficients. A la figura 2.9 es pot apreciar com es realitza la descomposició tenint en compte tres subbandes (wavelet de difusió en dos dimensions), en comptes de quatre (waveletes de dos dimensions per imatge).



**Fig. 2.9.** Dreta: arbre de descomposició (subespais definits i autovalors) per la wavelet de difusió en dos dimensions. Comparació amb la descomposició per una dimensió (esquerra) amb 10 autovalors. Las dimensions (n x m) en el cantó dret com a suport pels autovalors ( $\lambda$ ).

En els primers experiments amb DW en 2D es van estudiar més de 20000 matrius de teletrànsit tant d'Abilene com de Geant. Els dos objectius principals van ser veure com la difusió afecta a una matriu de teletrànsit amb la intenció d'analitzar la multiresolució subsegüent (comprovant la capacitat de reconstrucció, repetint el procés invers) i veure la compressibilitat obtinguda amb aquest mètode. Es va utilitzar com a operador el "random walk" sense ponderar normalitzat, i la precisió va ser de  $\epsilon=10^{-7}$ . Es pot apreciar la representació MRA d'una TM d'Abilene a la figura 2.11. Las aproximacions són els subespais  $VV$  (anomenats com a  $V$ , mentre que els detalls  $W$  són la suma

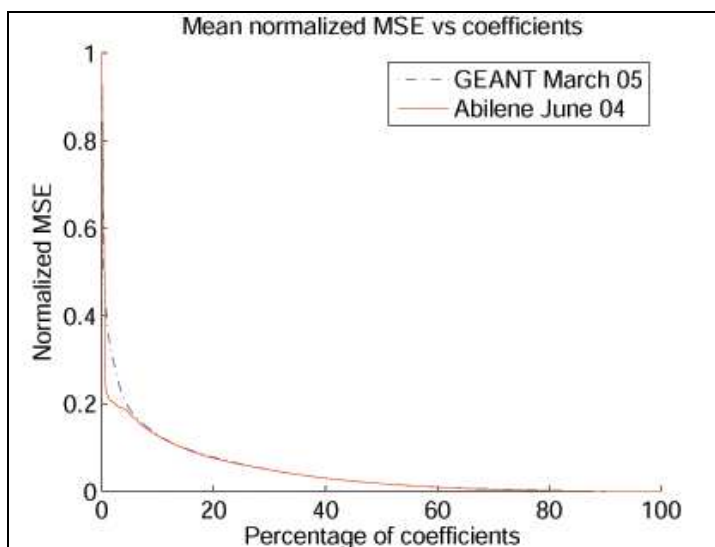
de WW, VW i WV). En els dos primers passos de la descomposició ( $V_1$  i  $V_2$ ) la reconstrucció és exactament la TM original degut a que  $W_1$  i  $W_2$  no tenen autovalors, per aquest motiu no han estat incloses aquestes imatges a la figura. L'efecte de les freqüències baixes es pot apreciar fàcilment a l'hora d'aplicar reiterativament l'operador a les aproximacions, com també l'efecte de les altes freqüències en els detalls.

Per la comprovació de la compressibilitat de la TM es van realitzar diferents tests agafant grans períodes de grups de dades (des de mig mes a la totalitat d'un, tant grans com permetien els "datasets" públics), amb l'objectiu de trobar quanta energia de la matriu original es mantenia concentrada en els primers coeficients. Es poden veure els resultats a la Taula 2.1.

**Taula 2.1.** Preservació de l'energia en els coeficients en dos grups de dades representatius (8640 TMs d'Abilene, i 2880 TMs per GÉANT).

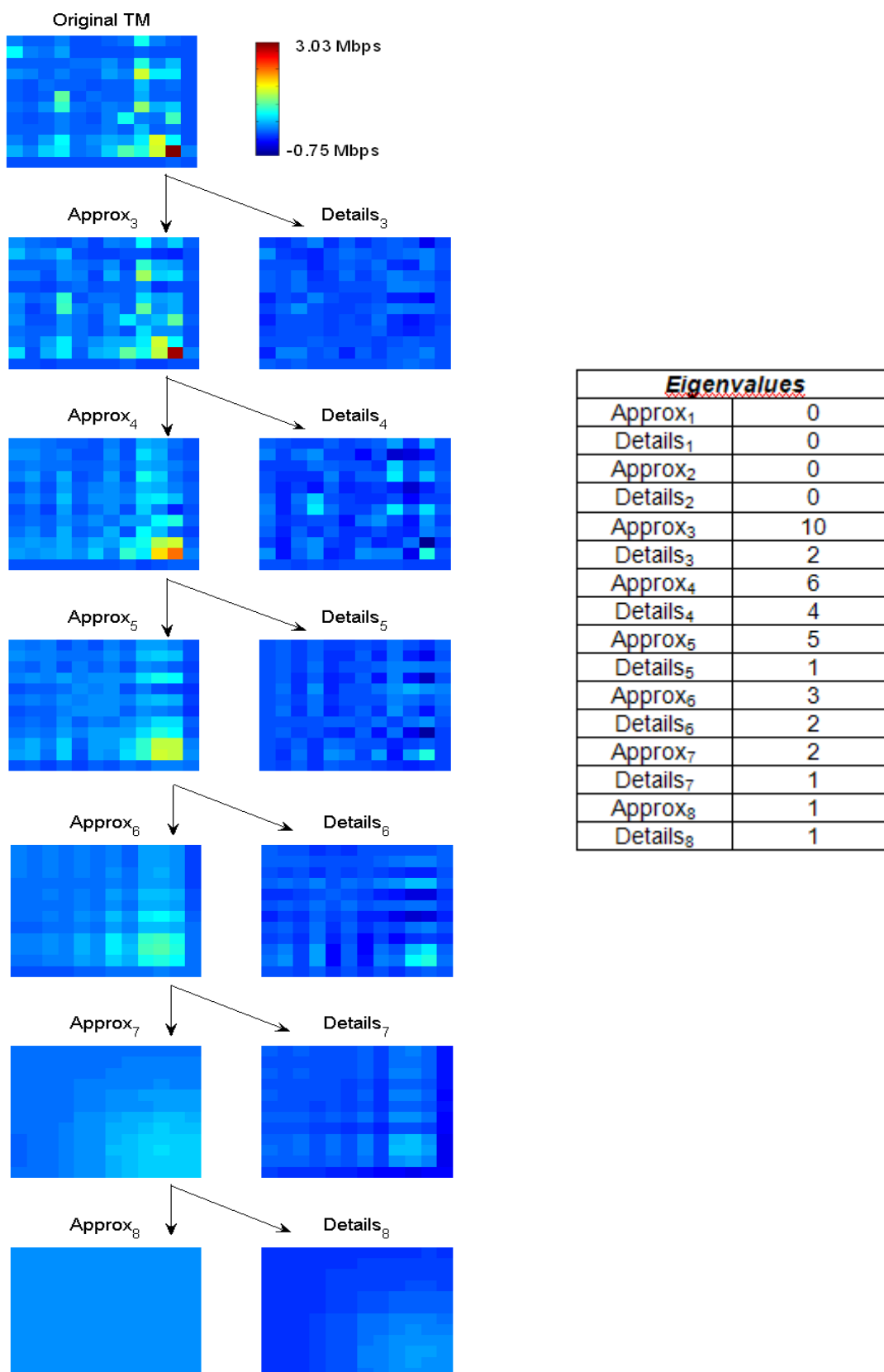
Set	Preserved Energy	
	80%	90%
Abilene, June 2004	4 coefs. (2.8%)	21 coefs. (14.6%)
GÉANT, March 2005	24 coefs. (4.5%)	75 coefs. (14.2%)

A la figura 2.11 es pot apreciar com la imatge s'esborra a mesura que avança en el procés de difusió (a cada pas s'analitzen subbandes més petites). Això és el que s'obté quan la difusió és realitzada sobre la xarxa.



**Fig. 2.10.** MSE normalitzat dependent del percentatge de coeficients de DW, per a un mes amb l'operador de difusió "random walk".

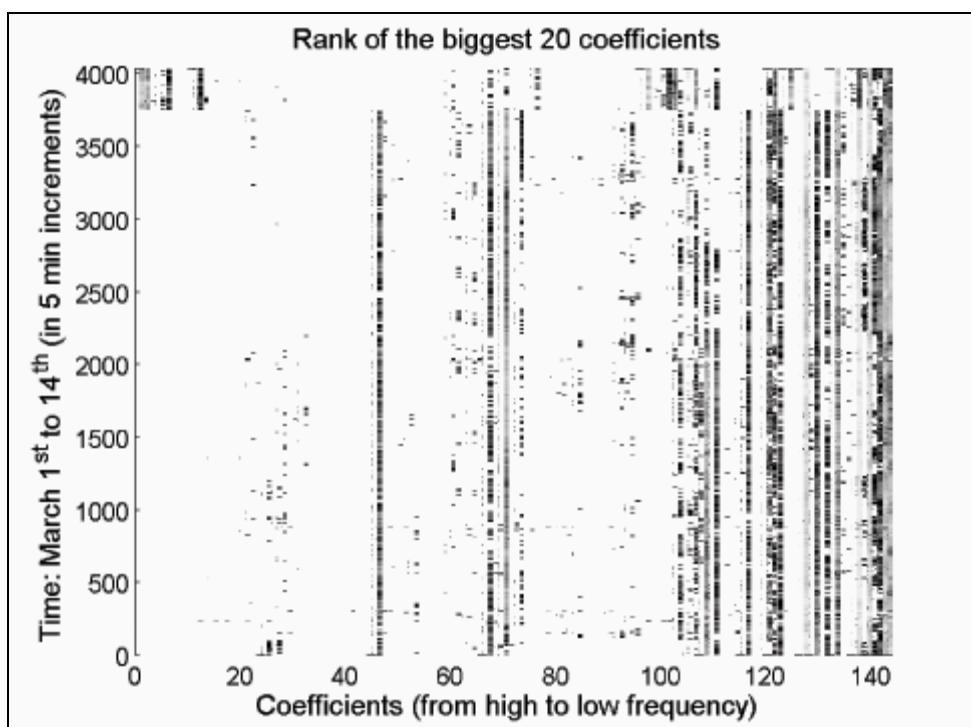




**Fig. 2.11.** Descomposició: aproximació i detalls dels subespais associats dels autovalors per l'anàlisi multiresolució del exemple mencionat.

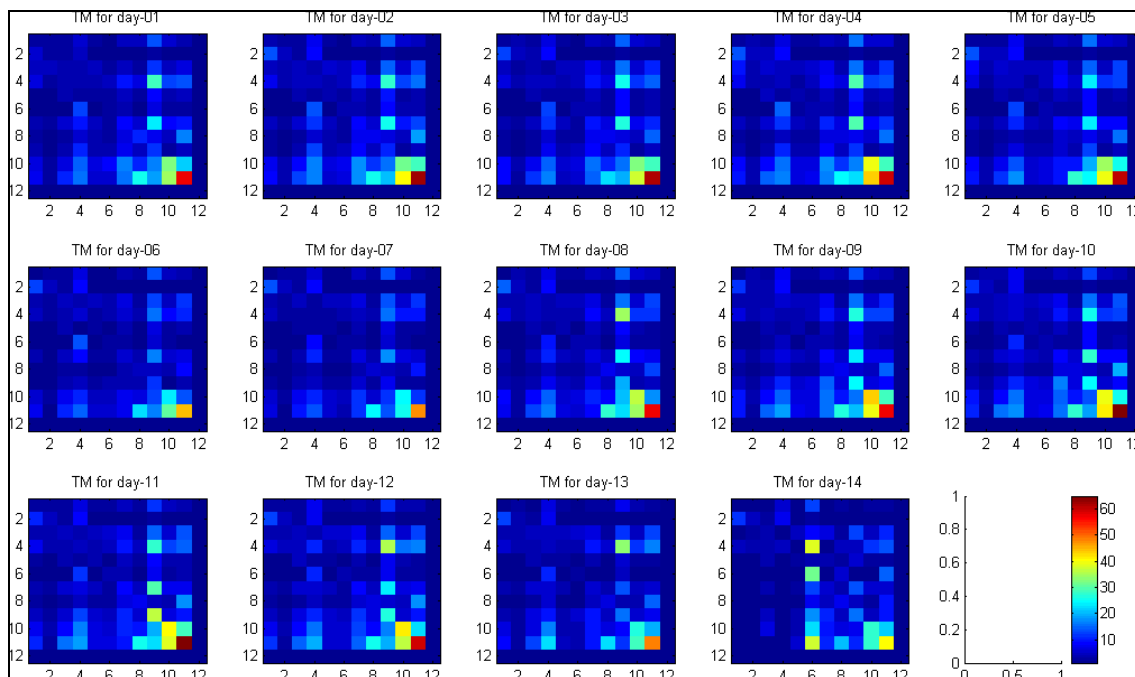
Els resultats obtinguts amb altres grups de dades temporals coincideixen amb els resultats de la taula anterior i confirmen la dispersió de la representació DW: de mitja, el 15% dels coeficients suposen el 90% de l'energia original de la TM. La figura 2.10 completa aquest anàlisi mostrant el Mean Square Error (MSE) de la TM reconstruïda tenint en compte el percentatge de coeficients utilitzats en la el procés de reconstrucció. Tot i les diferències entre les dades d'Abilene i Geant els resultats són molt semblants.

Un fet a notar és que cada matriu té la seva pròpia característica (com una firma), relacionada amb el valor dels coeficients (depenent de la seva contribució a l'energia), la qual serà persistent en el temps. A la figura 2.12 es mostra l'ordre relatiu (els coeficients més gran és el color més fosc) dels 20 coeficients més grans de 14 dies de la traça de Març 2004 de les TMs d'Abilene. A partir d'aproximadament la matriu 3700, es pot apreciar un canvi estructural clar en el 14<sup>e</sup> dia (com es pot veure a la Fig. 2.13, el node 6 comença a rebre una quantitat de teletrànsit que no rebia en els dies previs, mentre que el node 9 decreix la quantitat de teletrànsit rebut).



**Fig. 2.12.** Firma particular pels dies de l'1 al 14 de Març del 2004 d'Abilene.

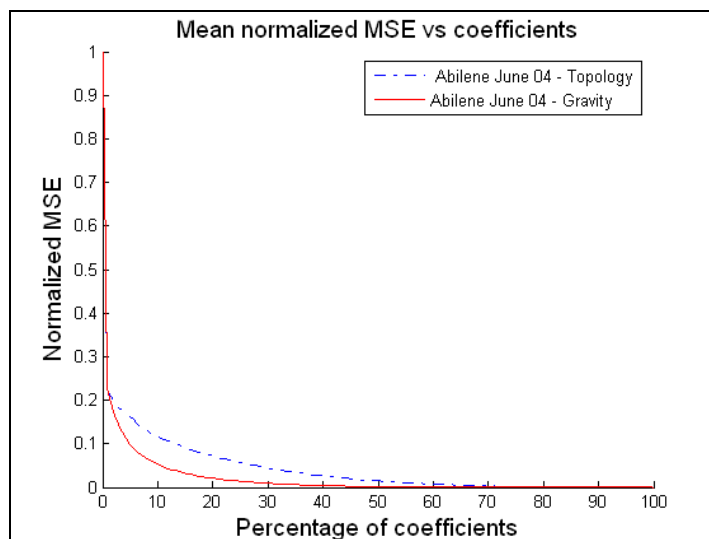
Aquests canvis poden ser mesurats només agafant uns pocs coeficients (si la seva energia és significant en la totalitat). Això està lluny de convertir-se amb un algoritme de detecció d'anomalies, però la transformada de DW sembla potencialment útil per aquest propòsit.



**Fig 2.13.** Valors de la mitja diària per els primers 14 dies de Març del 2004 d'Abilene. Es pot apreciar un canvi estructural el dia 14.

L'elecció de l'operador de difusió és un factor important en el procés, ja que defineix la base sobre la qual la funció estudiada (la TM, en el nostre cas) serà projectada. El resultat pel "random walk" ha estat presentat com exemple, ja que és l'operador més simple, però l'objectiu és trobar un bon operador. Per aquest propòsit, i de nou, com exemple, es mostra un altre operador de difusió basat en el model de gravetat. La idea és, malgrat treballar amb la topologia del graf, definir un altre graf en els enllaços siguin ponderats d'acord al model de gravetat, per exemple, els nodes que intercanvien gran quantitat de teletrànsit seran els "veïns" més propers i aquells que intercanviïn volums de teletrànsit baixos seran "llunyans". S'ha de recordar que el model de gravetat pot ser inferit pels comptadors SNMP i per tant ja està disponible per mesurament directe. A la figura 2.14 es mostra una comparació entre els resultats obtinguts per l'operador "random walk" i l'operador basat en el model de gravetat.

Aquests resultats són presentats a [10], i mitjançant el desenvolupament d'aquesta tesis s'han repetit, per pogué confirmar la seva fiabilitat. Els resultats mostren un millora en la compressibilitat quan es comparen amb les adjacències sense ponderar del "random walk".



**Fig. 2.14.** MSE normalitzat respecte el percentatge de coeficients pels dos operadors diferents sobre Abilene.

## CAPÍTOL 3. DADES ANALITZADES

### 3.1. Descripció de les dades

Les dades amb les que s'ha treballat han estat obtingudes de les xarxes acadèmiques americana, Abilene [32] i europea, Geant [33]. S'ha de tenir en compte que aquestes organitzacions eviten donar informació de la seva xarxa a un àmbit públic, per evitar que aquestes dades anessin a parar a usuaris maliciosos. Per tant, els investigadors han de demanar permisos especials per obtenir aquestes dades.

#### 3.1.1. Abilene

Abilene escull publicar les seves dades i el mapa de la xarxa, els enllaços de cada node als punts de presència (PoPs), però protegint aquestes dades perquè són obsoletes o amb desviacions temporals: els grups de dades amb els que s'ha treballat són suposadament del 2004, però no es pot assegurar.

#### 3.1.2. Geant

Geant oculta alguna informació a més de publicar dades antigues. Les seves dades són anònimes: no es poden conèixer quins nodes corresponen a cada PoPs, i a més es poden trobar mostres que no corresponen amb els noms que inicialment indiquen (per exemple, "IntraTM-2005-01-01-00-30", no han de ser obligatòriament dades de l'1 de gener del 2005 a les 00:30 hores).

Les mostres d'Abilene són recollides cada 5 minuts, per tenir un total de 12 matrius cada hora. En el cas de Geant només es té una matriu cada 15 minuts (4 per hora). Degut a això, és preferible treballar amb Abilene degut a que certs increments en els temps dels càlculs són proporcionals a la quantitat de nodes (per certs processos és millor treballar amb dotze matrius 12x12 per hora que no amb 4 matrius de 23x23 per hora, aquest últim cas és el de Geant).

S'ha d'indicar que les dades proporcionades per Abilene i Geant són incomplertes, algunes vegades només falta una matriu, però en altre poden haver forats d'hores o dies. En el cas d'Abilene, les dades proporcionades solen ser de grups de dies mitjanament llargs. En el cas de Geant les dades continues en dies o hores són més complexes d'obtenir. A l'hora de realitzar l'estudi aquest ha estat una de les circumstàncies que s'ha tingut en compte, ja que el fet de no tenir dades contínues, sobretot en el cas de Geant ha suposat descartar grups de dades que no donaven continuïtat als anàlisis i per tant una certa incongruència en els resultats. En el cas d'Abilene, al ser grups de dades més continus, a l'hora de realitzar certs càlculs ha permès visualitzar els cicles tant diaris com setmanals, fet que amb Geant, com s'ha comentat degut a la seva fraccionalitat, ha estat més complexa veure aquests cicles.

## 3.2. Utilització de les dades

Una vegada es van obtenir les dades, per poder treballar amb elles, es va tenir que estudiar la seva estructura. En un principi es va treballar amb les dades d'Abilene, que al ser més contínues permetien un tractament més simple que les de Geant. El projecte TOTEM [22], hagués permès mostrar les topologies d'ambdues xarxes i incorporar TM's per visualitzar comportaments d'aquestes, però no ha estat utilitzat. Per a realitzar els anàlisis numèrics s'ha utilitzat Matlab [21]. Disposem de funcions de Matlab que fan que tractament de les dades sigui relativament simple.

### 3.2.1. Anàlisi de les dades

Les matrius de teletrànsit són donades en format XML, tant les de Abilene com les de Geant.

Abilene:

```
<TrafficMatrixFile
xsi:noNamespaceSchemaLocation=http://totem.run.montefiore.ulg.ac.be/Schema/TrafficMatrix-v1\_2.xsd
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
<IntraTM ASID="11537">
<src id="ATLA-M5">
<dst id="ATLA-M5">
26.6666666666667
</dst>
<dst id="ATLAng">
522.208
</dst>
<dst id="CHINg">
1641.33866666667
</dst>
<dst id="DNVRng">
335.728
</dst>
<dst id="HSTNng">
413.032
</dst>
<dst id="IPLSng">
489.87466666667
</dst>
<dst id="KSCYng">
365.07733333333
</dst>
```

Geant:

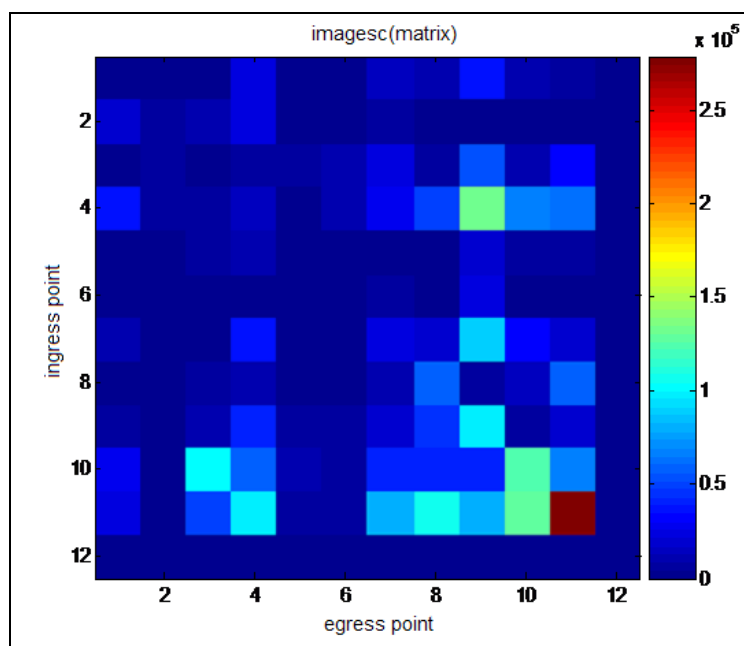
```
<TrafficMatrixFile
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:noNamespaceSchemaLocation=http://totem.run.montefiore.ulg.ac.be/Schema/TrafficMatrix-v1\_2.xsd >
<info>
<title>
Intradomain traffic matrix generated by geant-TM-all.pl
</title>
<date>
2005-01-01T00:30:00
</date>
<duration>
900
</duration>
<author>
Steve Uhlig
</author>
<units>
<unit type="bandwidth" value="kpbs"/>
</units>
</info>
<IntraTM ASID="20965">
<src id="12">
<dst id="12">
396708.2311
</dst>
```

**Fig.3.1.** Exemples parcials de dues matrius de teletrànsit . Esquerra: Abilene. Dreta: GEANT.

### 3.2.2. TM's a Matlab

Matlab és una aplicació matemàtica desenvolupada per Mathworks. Ens permet analitzar estadísticament les matrius, Matlab és una eina que ens ofereix una gran quantitat de funcionalitats relacionades amb una gran diversitat de camps matemàtics.

Una característica que ens aporta Matlab és la possibilitat de visualitzar les TMs. Amb aquest propòsit, s'ha desenvolupat codi per passar les dades dels arxius XML a Matlab i es puguin visualitzar. Si es treballa sobre les matrius directament, és a dir, amb els valors numèrics, és molt difícil veure les diferències i detectar cap mena de patró o anomalia. Tot i les diverses formes que té Matlab per representar matrius de dades, en aquest estudi només s'ha utilitzat la forma de representació **imagesc**. A la figura 3.2 es pot apreciar el tipus de gràfic que aporta Matlab amb l'ordre `imagesc`. La matriu representada és TM-2004-06-01-1300 (1 de Juny a la 13:00).



**Fig.3.2.** Representació de la TM-2004-06-01-1300 d'Abilene, sota el forma `imagesc` de Matlab.

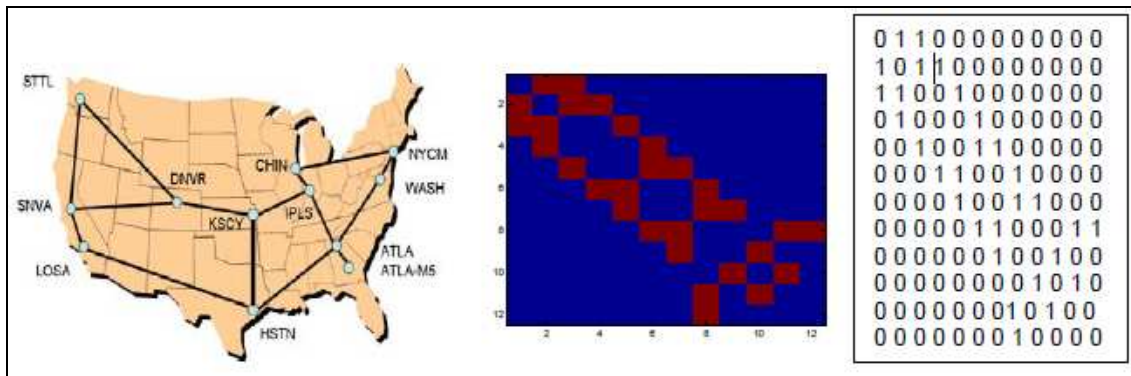
Aquest tipus de gràfic ens permet obtenir una representació visual dels paràmetres de la matriu de teletrànsit. L'associació gràfica dels colors va lligada al valor numèric de cada posició de la matriu. Per tant, en el cas de visualitzar un bloc de matrius consecutives es podrien detectar patrons de comportament o anomalies d'una forma molt senzilla si els colors, i per tant els valors relacionats a aquests, variessin.

### 3.2.3. Matrius d'adjacència i Laplacianes a Matlab

Un altre tipus de matrius que es visualitzaran a l'estudi són les matrius d'adjacència. Aquestes matrius indiquen quins nodes tenen enllaços directes entre ells. Sempre que la relació d'enllaç entre un node i un altre sigui directa aquest tindrà un valor igual a 1 a la posició de la matriu que relacioni aquests nodes.

A mode d'exemple, si disposem d'una matriu  $B$ , la matriu d'adjacència es representaria de la següent manera:  $b_{ij} = 1$  quant existeix un enllaç entre els nodes  $i$  i  $j$ ,  $b_{ij} = 0$  si no existeix enllaç directe entre els nodes  $i$  i  $j$ . D'aquesta

manera la matriu binària ens indica la relació de veïnatge entre els nodes a la xarxa o graf i és molt útil quant es prova l'operador de difusió.



**Fig.3.3:** Relació entre la xarxa Abilene (graf) i la seva matriu d'adjacència.

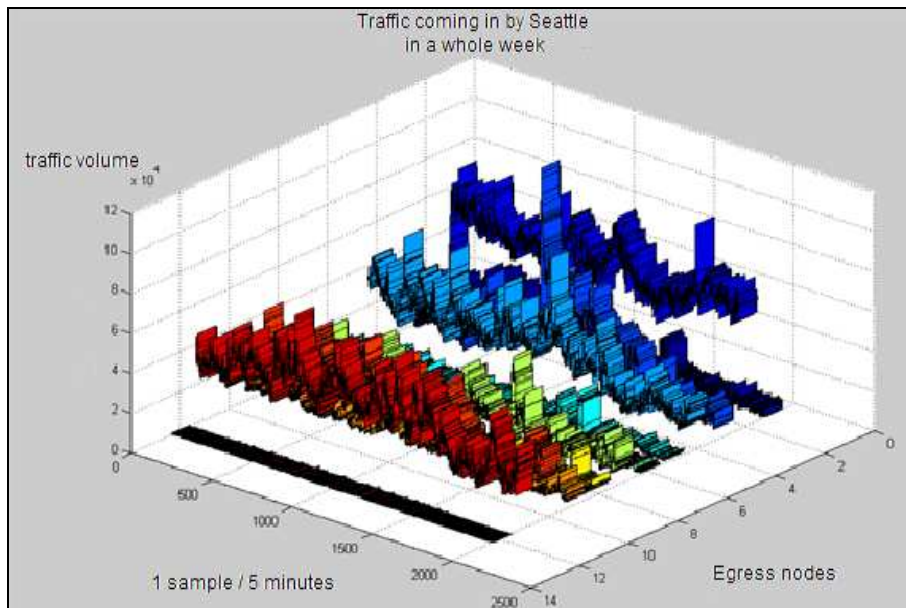
A la figura 3.3 es pot trobar la representació de la matriu d'adjacència de la xarxa Abilene i els seus valors. La analogia és visible: els pixels vermells són els que tenen valor 1 i els pixels blaus són els que tenen valor 0. Cada valor  $[i,j]$  representa una combinació dels dos nodes: el primer 0 representa l'adjacència de Seattle a Seattle, (aquest és zero per que ell mateix no pot ser el seu propi veí). Si s'observa la primera línia, es trobarà els nodes veïns de Seattle en aquest ordre: Sunnyvale i Denver (la distribució es pot veure a la figura 1.1, Abilene). Es pot apreciar que en aquest cas la diagonal de la matriu d'adjacència és zero (la diagonal representa las adjacències de cada node amb ell mateix) encara que en el cas de l'estudi també es considera el teletrànsit que va d'un node cap a ell mateix (i per tant haurà de ser afegit a la matriu identitat).

Existeix una variació d'aquesta matriu que és la *matriu Laplaciana* o també anomenada *matriu d'admitància* de l'adjacència. La matriu Laplaciana  $L$ , és definida sobre un graf de la següent manera:  $l_{ij} = -1$  si  $i$  i  $j$  són veïns (vèrtex  $v_i$  i  $v_j$  enllaçats per una vora),  $l_{ij} = deg(v_i)$  només quan  $j = i$ , i  $l_{ij} = 0$  en qualsevol altre cas. La matriu Laplaciana mostra el veïnatge dels nodes en el graf: això està íntimament relacionat amb el random walk, de manera que també són importants en la definició de l'operador de difusió. En termes generals, quan es realitza una MRA sobre un graf, la relació amb el veïnat es pot considerar un bon punt de partida.



## CAPÍTOL 4. CORRELACIÓ TEMPORAL

L'estudi s'inicia a partir de la hipòtesis que existeix una correlació temporal entre la quantitat de teletrànsit transportada en una ruta en un temps  $t = n$  i  $t = n+1$ , per tant, en mostres consecutives. Seguint aquesta línia de recerca l'objectiu es fer un operador basat en aquesta autocorrelació apreciament la posterior difusió i compressibilitat.



**Fig.3.3:** Volum de teletrànsit entrant en el node de Seattle en el transcurs d'una setmana. Es pot apreciar una regularitat entre els set gràfics.

### 4.1. Coeficient de correlació

L'objectiu d'aquesta primera part de l'estudi va ser obtenir un valor que relacionés les diferents matrius de teletrànsit entre elles. El paràmetre que es va utilitzar per relaciona les diferents matrius que es disposava de les xarxes a estudi va ser el volum de teletrànsit entre els diferents nodes.

S'ha de tenir en compte a l'hora de realitzar l'anàlisi que les dues xarxes estudiades tenen una diferència de comportament fonamental. El comportament diferencial es troba en l'ús per part del usuaris de cadascuna de les xarxes. Abilene és una xarxa amb un volum de teletrànsit elevat, d'un ús molt freqüent entre els usuaris que en fan ús, volum constant de fluxos de teletrànsit. Pel que fa a Geant en el moment de la recollida de las dades, l'any 2005, era una xarxa d'un ús molt més irregular, un teletrànsit de dades esporàdic, a ràfegues.

Aquesta característica diferencial de les dues xarxes va generar que a l'hora de calcular el coeficient de correlació entre les diferents matrius aquest paràmetre fos calculat també de forma diferenciada.

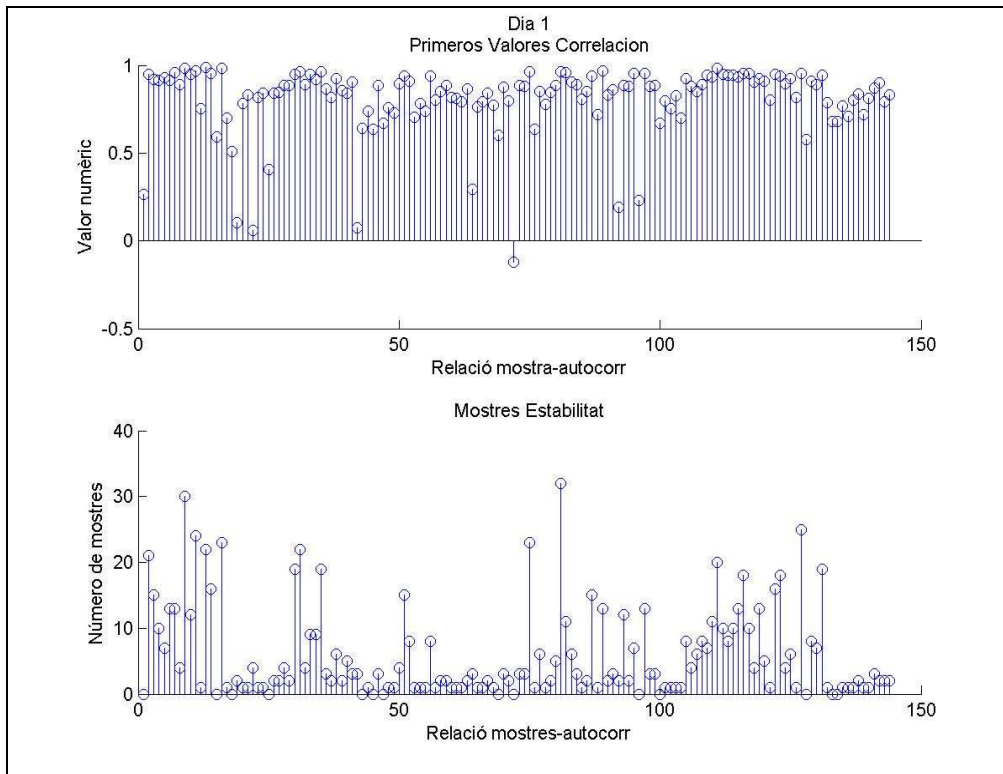
En el cas de les dues xarxes el coeficient que es va buscar va ser el valor d'autocorrelació diari mig del teletrànsit per a cada ruta, però com s'ha comentat amb anterioritat en el cas d'Abilene, degut al comportament del teletrànsit entre rutes, el procés va ser més elaborat que no pas en el cas de Geant.

#### **4.1.1. Correlació Abilene**

L'estudi es va realitzar obtenint els valors d'autocorrelació de cada una de les rutes per a cada un dels 167 dies que es disposava per fer l'estudi. En un principi aquest valor d'autocorrelació mig va ser de 0,750852, però l'estudi de les autocorrelacions de cadascuna de les rutes va desvelar que existien algunes que, ja fos en moments puntuals o d'una forma continuada, donaven valors d'autocorrelació molt baixos, inclús negatius. Es va determinar que aquells valors d'autocorrelació que estiguessin per sota de 0,7 serien considerats inestables. Tenint en compte aquest fet, es va procedir a realitzar un anàlisi de les rutes en concret que presentaven valors inferiors al considerat d'estabilitat.

A la Figura 4.1 es pot apreciar al gràfic superior els valors d'autocorrelació mig per un dia per a cadascuna de les rutes, i el gràfic inferior la quantitat de valors d'autocorrelació considerats estables per a cada ruta en valors mitjos d'autocorrelacions per hora del dia estudiat.

Com s'ha explicat anteriorment, fent l'estudi d'inestabilitat es va apreciar que existien rutes que eren considerades inestables de forma puntual, però n'hi havia que ho eren de forma constant. Per determinar com d'inestables eren cadascuna de les rutes es va procedir a fer un càlcul de dies d'inestabilitat mostrat per cadascuna de les 144 rutes que disposa Abilene. Aquest càlcul va mostrar que hi havia rutes que es mantenien inestables al llarg del temps i es va determinar d'eliminar del càlcul del coeficient d'autocorrelació global el valors d'autocorrelació mitjos mostrats per les rutes considerades inestables. Aquest pas es va determinar pel fet que es va considerar que aquests valors d'autocorrelació falsejaven el valor mig global d'autocorrelació. Per considerar una ruta inestable es va agafar com a criteri el determinar com a tal aquelles rutes que es mantenien inestables aproximadament el 50% dels dies totals que es disposava per l'estudi, es a dir sobre els 84 del 167 dies



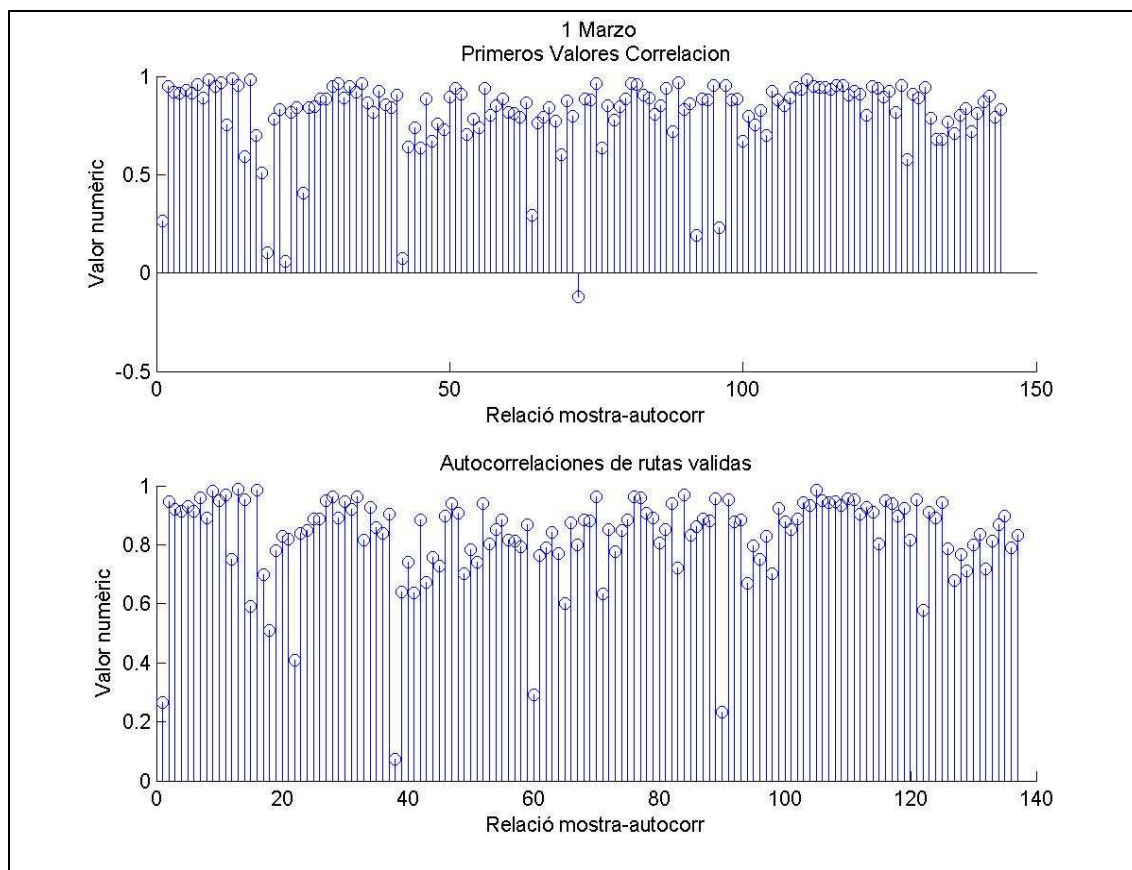
**Fig. 4.1.** Gràfic superior: Valors autocorrelació mitjos per a cada ruta pel dia 1 de març del 2004. Gràfic inferior: Quantitat de valors d'autocorrelació estables per mitges horàries de cada ruta pel dia 1 de març del 2004.

El número total de rutes d'Abilene són 144, les rutes descartades pel càlcul de valor d'autocorrelació global van ser 7 :

1. Ruta SNVA-IPLS, dies d'inestabilitat 145.
2. Ruta SNVA-NYCM, dies d'inestabilitat 117.
3. Ruta SNVA-ATLA-M5, dies d'inestabilitat 96.
4. Ruta DNVR-ATLA-M5, dies d'inestabilitat 81.
5. Ruta HSTN-ATLA-M5, dies d'inestabilitat 140.
6. Ruta ATLA-ATLA, dies d'inestabilitat 125.
7. Ruta ATLA-M5-SNVA, dies d'inestabilitat 102.

De les set rutes descartades per realitzar el càlcul del valor d'autocorrelació es pot identificar que en quatre d'aquestes intervé el node ATLA-M5. Aquest node, com es pot veure a la figura 1.1, la seva comunicació amb el resta de la xarxa depèn únicament del node ATLA, aquest fet genera que el teletrànsit generat o amb destinació a ATLA-M5 sigui molt poc (i per tant més sensible a variacions). Tot i això fa que el trànsit d'aquest enllaç sigui molt variable.

A la Figura 4.2 es pot apreciar al gràfic superior els valors mitjos d'autocorrelació de cadascuna de les rutes per un dia, i en el gràfic inferior es mostren els valors d'autocorrelació únicament de les rutes que s'han considerat estables.



**Fig. 4.2.** Gràfic superior: Valors autocorrelació mitjos per a cada ruta pel dia 1 de març del 2004. Gràfic inferior: Valors d'autocorrelació de cada ruta estable per mitges horàries de cada ruta pel dia 1 de març del 2004.

Com es pot veure a la Figura 4.2 , el número de rutes inicials són les 144 totals d'Abilene, però les utilitzades pel càlcul del valor d'autocorrelació son 137, com es veu al gràfic inferior de la mateixa figura.

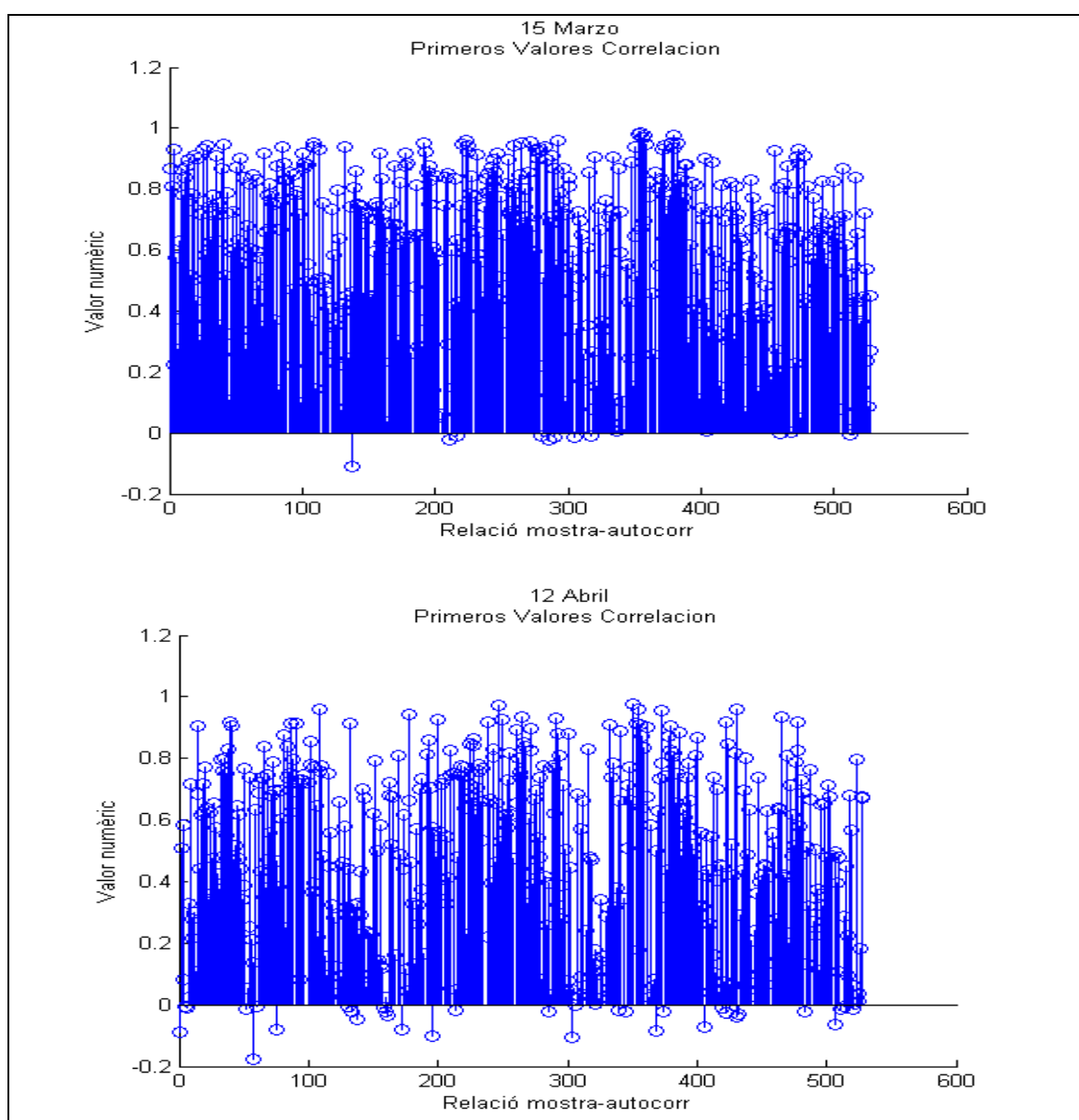
Una vegada es va prendre la decisió de descartar les rutes inestables es va tornar a realitzar el càlcul del coeficient d'autocorrelació donant com a resultat un valor sensiblement superior que l'inicial, exactament 0,773337. D'aquest valor, pels càlculs posteriors de l'estudi, van ser descartats el quatre últims decimals, obtenint 0,77 com a coeficient d'autocorrelació global per l'anàlisi espai-temporal multiresolució de matrius de teletrànsit.

#### 4.1.2. Correlació Geant

Pel que fa a Geant l'estudi de l'autocorrelació va tenir una dificultat afegida, com s'ha comentat anteriorment, la falta de continuïtat en les mostres que es disposaven. Per aquest fet es van tenir que descartar dies sencers pel càlcul de la autocorrelació, ja que els vuits de matrius intercalats en mig del que en un principi semblaven períodes llargs d'anàlisi, generaven una distorsió del resultat, ja que les matrius no eren consecutives i per tant la correlació entre matrius no era real. Finalment es van pogué estudiar 106 dies sencers, amb els seus corresponents valors d'autocorrelacions per cada ruta.

El coeficient de correlació entre les diferents rutes va desvelar una inestabilitat molt més elevada que en el cas d'Abilene. Aquesta inestabilitat, com s'ha comentat anteriorment, ve determinada pel comportament d'us de la xarxa. Geant presenta un comportament d'us molt esporàdic, i a més, les rutes de teletrànsit són molt constants, es a dir, els fluxos de dades solen ser entre rutes concretes. Tenint en comte aquests fets es va determinar el no descartar cap valor d'autocorrelació ni cap ruta, a diferència de l'estudi realitzat amb Abilene, a l'hora de trobar el valor de correlació entre matrius.

A la figura 4.3 es pot apreciar la inestabilitat de les rutes comentada. S'ha de tenir en compte que els dies que apareixen als gràfics no corresponen a la realitat, ja que les dades de Geant són parcialment confidencials, el dia a que fa referència el gràfic és a l'indicat en els arxius XML que contenen la matriu de teletrànsit avaluada.



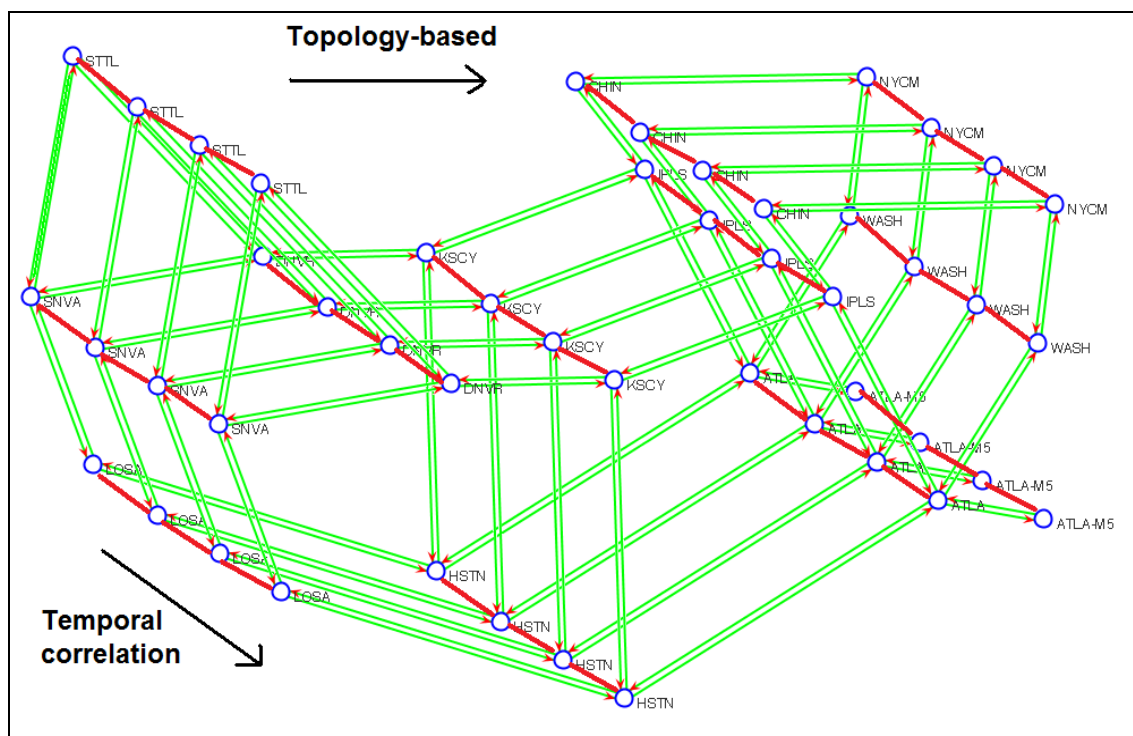
**Fig 4.3.** Superior i inferior: correlació de les 529 rutes de Geant en diferents dies.

Una vegada realitzat aquest anàlisi sobre les rutes de Geant el valor d'autocorrelació va ser 0.55671, com es pot apreciar molt més baix que l'obtingut a Abilene degut, com s'ha explicat, a la inestabilitat mostrada a la xarxa. En aquest cas es van mantenir tots els decimals del valor obtingut.

## 4.2. Anàlisi espai-temporal

Una vegada es tenen les matrius de teletrànsit reals el repte és relacionar cadascuna d'aquestes matrius amb les altres. Generant un espai tridimensional, és a dir, amb les matrius obtenim una relació espacial en 2D, nodes origen i final, la relació que s'ha buscat és la temporal de cada matriu amb las següents i anteriors.

La figura 4.4 mostra la idea de relacionar temporalment las matrius de teletrànsit sobre la topologia de la xarxa origen de les matrius estudiades, en aquest cas Abilene.



**Fig 4.4:** Representació de la topologia d'Abilene en el temps.

Com s'ha explicat anteriorment, el càlcul del valor d'autocorrelació es va realitzar per obtenir un valor matemàtic que ens relaciones temporalment les matrius estudiades.

Una vegada obtingut el valor d'autocorrelació i conjuntament amb el model de gravetat es va realitzar l'anàlisi espai-temporal. L'anàlisi es va realitzar amb cinc mètodes diferents. Van ser quatre mètodes on es va introduir el valor d'autocorrelació d'una o altre forma i un mètode on no es va introduir el valor.

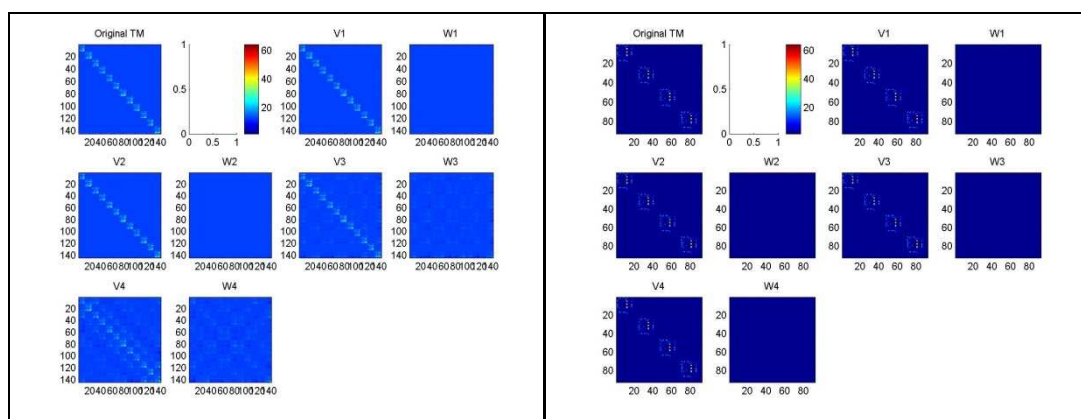




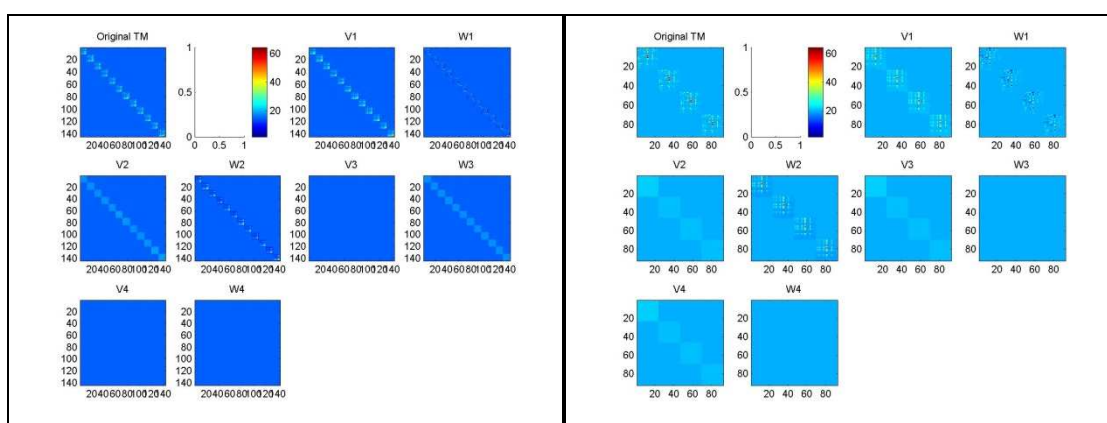




coeficient de correlació es va descartar després d'obtenir els resultats, ja que eren pitjors als obtinguts amb els mètodes mostrats.

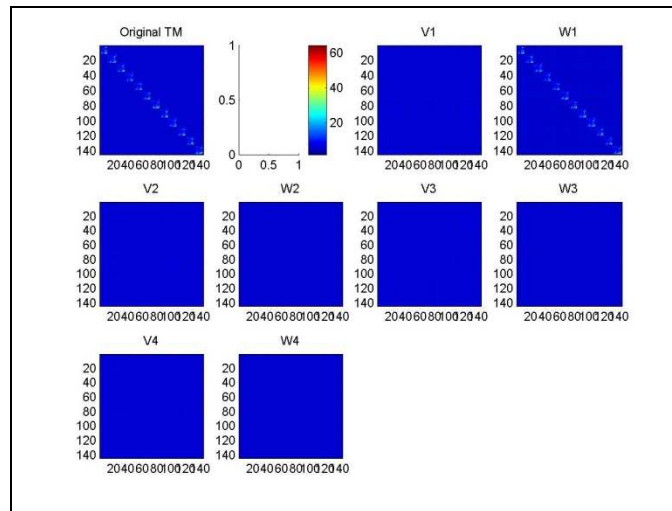


**Fig. 4.7.** Mètode 3 d'estudi de les aproximacions i detalls de l'anàlisi multi-resolució de matrius. Esquerra: Xarxa Abilene, matrius TM-2004-07-05-1800 a TM-2004-07-05-1855. Dreta: Xarxa Geant, matrius IntraTM-2005-01-06-11-00 a IntraTM-2005-01-06-11-45.



**Fig. 4.8.** Mètode 4 d'estudi de les aproximacions i detalls de l'anàlisi multi-resolució de matrius. Esquerra: Xarxa Abilene, matrius TM-2004-07-05-1800 a TM-2004-07-05-1855. Dreta: Xarxa Geant, matrius IntraTM-2005-01-06-11-00 a IntraTM-2005-01-06-11-45.

Es va realitzar un últim mètode de càlcul en el que no s'incorporava en cap moment el valor d'autocorrelació, per poder valorar els resultats i observar si realment el coeficient que s'estava utilitzant suposava un avanç. Com posteriorment es podrà observar, els resultats donen crèdit a la utilització d'aquest valor d'autocorrelació com un pas endavant en l'anàlisi espai-temporal de matrius de teletrànsit.



**Fig. 4.9.** Aproximacions i detalls de l'anàlisi multi-resolució de les matrius TM-2004-07-05-1800 a TM-2004-07-05-1855. Mètode 5

Com s'ha comentat anteriorment els mètodes 1,2 i 5 van ser descartats degut a que els mètodes 3 i 4 van generar millors resultats. La figura 4.9 mostra l'anàlisi multi-resolució amb el mètode 5 a mode de referència per avaluar l'efecte del coeficient de correlació sobre el model aplicat, en el nostre cas una variació del model de gravetat.

## CAPÍTOL 5. RESULTATS

### 5.1. Abilene

La compressibilitat del resultats s'ha realitzat des de dos punts de vista diferenciats. El primer ordre de resultats es van basar en el càlcul del error quadràtic normalitzat, MSE, en funció del percentatge de coeficients de la DW.

El segon resultats es van obtenir la relació entre els coeficients de la matriu i la quantitat d'energia associada aquests coeficients. Els coeficients obtinguts, estan relacionats amb el valor d'  $\epsilon$ .

#### 5.1.1. MSE

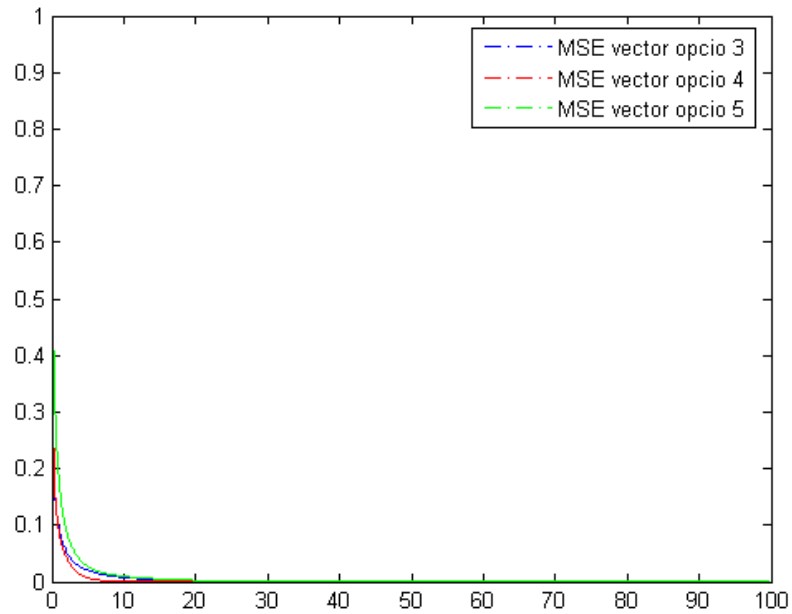
El càlcul del error quadràtic normalitzat ens permet saber el percentatge de coeficients necessaris per reconstruir la matriu original una vegada fet l'anàlisi multiresolució. Quants menys coeficients siguin necessaris per la reconstrucció, més bo serà el resultat.

S'han diferenciat els resultats de les dues xarxes per un problema computacional. Degut a que el mostreig de les dades que es disposen d'Abilene són d'una matriu cada cinc minuts, els càlculs per hora es multiplicant per dotze en cas de realitzar un anàlisi horari. En el cas de Geant el mostreig és d'una matriu cada quinze minuts, això suposa de quatre matrius per hora.

El millor mode de visualitzar els resultats és comparant aquest amb els que fins ara s'havien aconseguit en altres estudis. A la figura 2.14 es pot apreciar l'error quadràtic mig normalitzat en funció del percentatge de coeficients de la DW en estudis anteriors amb l'aplicació de dos models diferenciats.

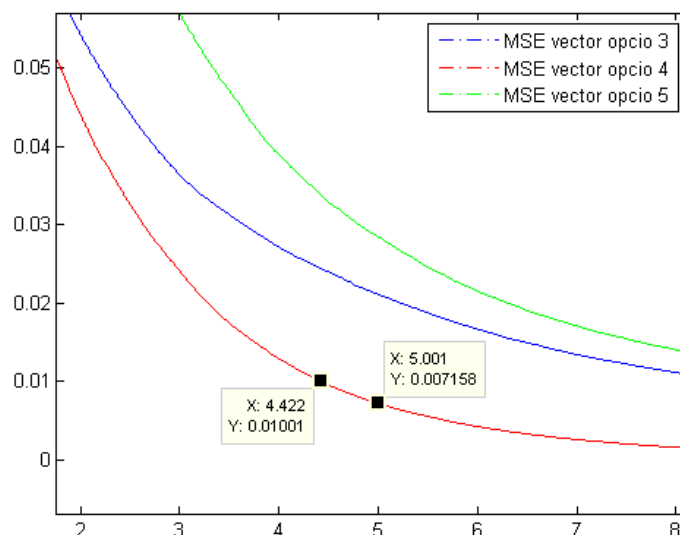
Amb Abilene es va realitzar un primer estudi sobre una hora, 12 matrius, concretament de les 18:00 a les 18:55 del dia 5 de Juliol del 2004, amb la que es va realitzar els càlculs amb els diferents mètodes. En el gràfic 5.1 es pot apreciar l'error quadràtic mig normalitzat en funció del percentatge de coeficients de la DW en els dos mètodes que van donar els millors resultats i el mètode sense la incorporació del coeficient de correlació, per apreciar la incidència d'aquest a l'estudi.

D'aquests gràfics es poden extreure dues conclusions directes. El primer, referent a la comparació amb els gràfics de la figura 2.14, és que els resultats són considerablement millors en l'estudi actual. La segona de les conclusions és la clara incidència de la incorporació del valor de correlació en l'estudi. Com es pot veure a la figura 5.1, el mètode que no incorpora el coeficient de correlació en la "supermatriu", opció d'estudi 5, tot i millorar els resultats d'anteriors estudis és el pitjor dels tres mostrats. La millor de les tres opcions és aquella que incorpora el coeficient de correlació en totes les posicions adjacents a la diagonal formada per les matrius probabilístiques, opció 4 d'estudi.



**FIG. 5.1.** Xarxa Abilene, 12 matrius, de la TM-2004-07-05-1800 a TM-2004-07-05-1855. Error quadràtic mig normalitzat en funció del percentatge de coeficients de la DW, amb els tres mètodes d'incorporació del coeficient de correlació estudiats.

A la figura 5.2 es mostra una ampliació dels resultats, on es pot apreciar clarament que amb el mètode 4 de càlcul, amb un 5% dels percentatge de coeficients s'aconsegueix la reconstrucció de la matriu original en pràcticament la seva totalitat.



**FIG. 5.2.** Ampliació de l'error quadràtic mig normalitzat en funció del percentatge de coeficients de la DW, amb els tres mètodes d'incorporació del coeficient de correlació estudiats.

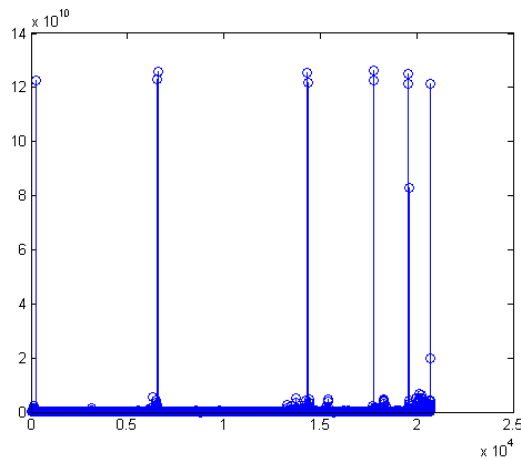
### 5.1.2. Distribució de l'energia entre els coeficients

La segona valoració per analitzar els resultats va ser calcular la relació entre els coeficients de la matriu i la quantitat d'energia associada aquests coeficients.

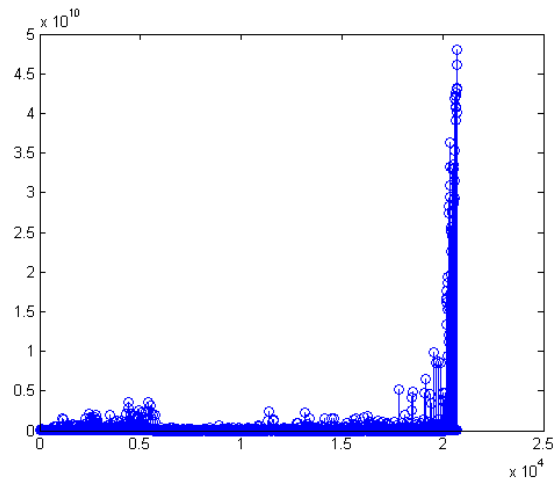
Els resultats en aquest cas mostraran quin percentatge de coeficients calculats contenen una major energia de la matriu, sent un valor més positiu aquells que en pocs coeficients concentrin més energia.

En un primer terme podrem apreciar els resultats per un valor d' $\varepsilon = 10^{-8}$  pels tres mètodes d'anàlisi estudiats, d'aquesta manera es podran veure les diferències entre ells. En segona instància es realitzarà la comparació del millor dels resultats dels tres mètodes d'anàlisi, opció 4, amb el mateix mètode però amb el valor d' $\varepsilon = 10^{-15}$ .

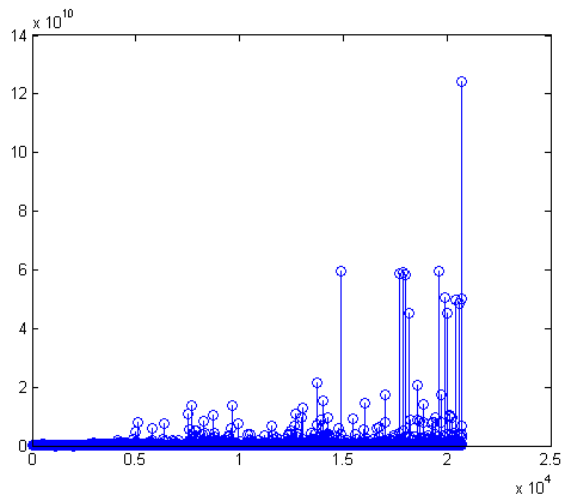
A les figures 5.3, 5.4 i 5.5 es pot observar l'energia dels coeficients obtinguts utilitzant les opcions 3, 4 i 5 d'anàlisi amb un  $\varepsilon = 10^{-8}$ . Una vegada més a l'opció 4 és fàcilment apreciable la concentració de coeficients en el gràfic respecte als altres dos mètodes d'estudi. Les altres opcions mostren una dispersió molt més elevada.



**FIG. 5.3.** Eix x: coeficients. Eix y: valor dels coeficients. Mètode d'estudi 3.

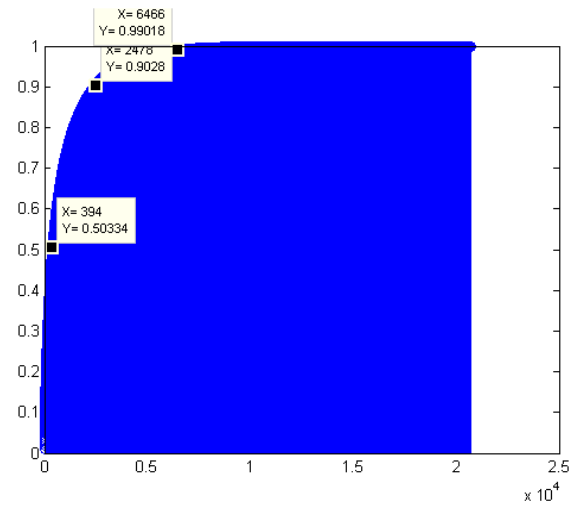


**FIG. 5.4.** Eix x: coeficients. Eix y: valor dels coeficients. Mètode d'estudi 4.

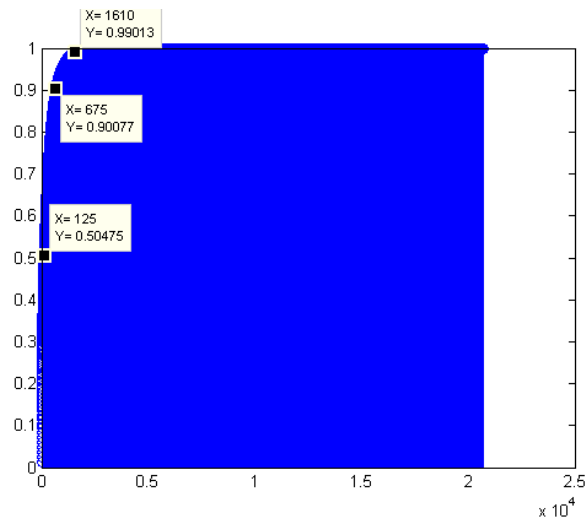


**FIG. 5.5.** Eix x: coeficients. Eix y: valor dels coeficients. Mètode d'estudi 5.

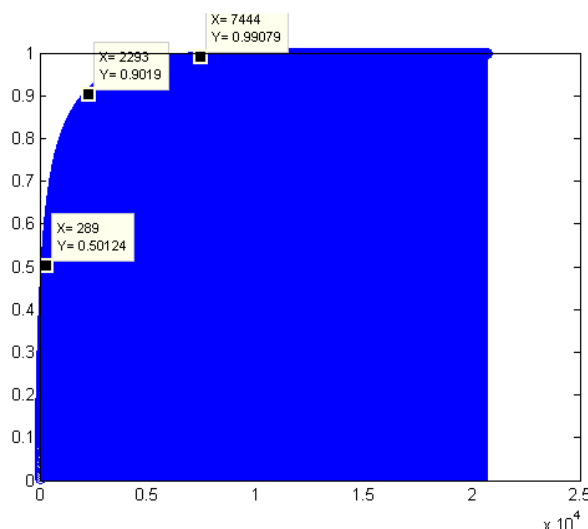
Una vegada obtinguts els coeficients per reflexa els resultats s'ha calculat l'energia acumulada associada a aquests.



**FIG. 5.6.** Energia acumulada respecte els coeficients. Eix x: quantitat de coeficients. Eix y: valor percentual de la quantitat d'energia. Mètode d'estudi 3.



**FIG. 5.7.** Energia acumulada respecte els coeficients. Eix x: quantitat de coeficients. Eix y: valor percentual de la quantitat d'energia. Mètode d'estudi 4.



**FIG. 5.8.** Energia acumulada respecte els coeficients. Eix x: quantitat de coeficients. Eix y: valor percentual de la quantitat d'energia. Mètode d'estudi 5.

Segons els gràfics de les figures 5.6, 5.7 i 5.8, podem extreure dues conclusions. La primera és que el mètode 4, com s'ha demostrat al llarg de l'estudi és el que millor resultats aporta. En aquests gràfics el mètode 4 mostra clarament que en l'energia de las TMs estudiades es reflecteix en un percentatge molt menor de coeficients que en els altres dos mètodes.

A la següent taula queda reflectit aquest fenomen:

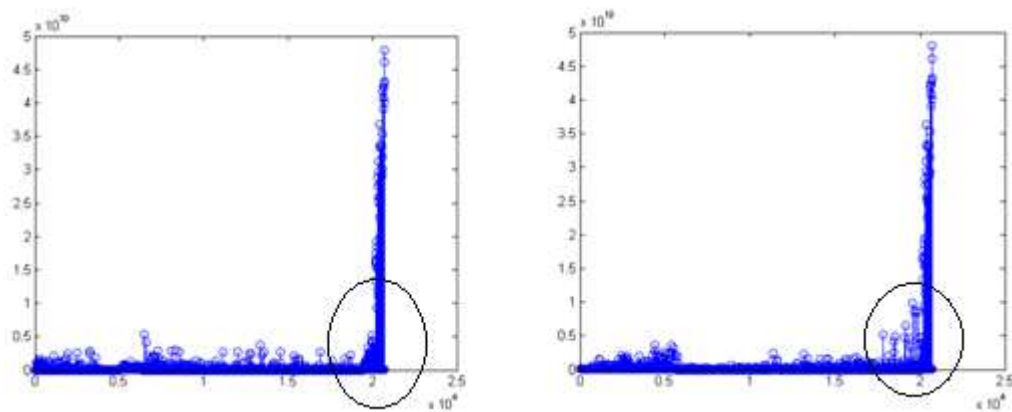
% Energia	% coeficients Mètode 3	% coeficients Mètode 4	% coeficients Mètode 5
50%	1,885%	0,602 %	1,393 %
90%	11,950 %	3,255 %	11,058 %
99%	31,182 %	7,764 %	35,8989 %

**Taula 5.1.** Comparació percentual de quantitat de coeficients respecte l'energia amb els tres mètodes d'estudi.

Una altra conclusió que es pot extreure del gràfics referents a l'energia és importància en la forma en que s'introdueix el valor d'autocorrelació a l'operador de difusió. La taula 5.1 mostra que en el mètode 5, aquell que no incorpora el coeficient d'autocorrelació en el càlcul, ofereix millors resultats per percentatges del 50 i 90 per cent que no pas el mètode 3, el qual incorpora el coeficient d'autocorrelació.

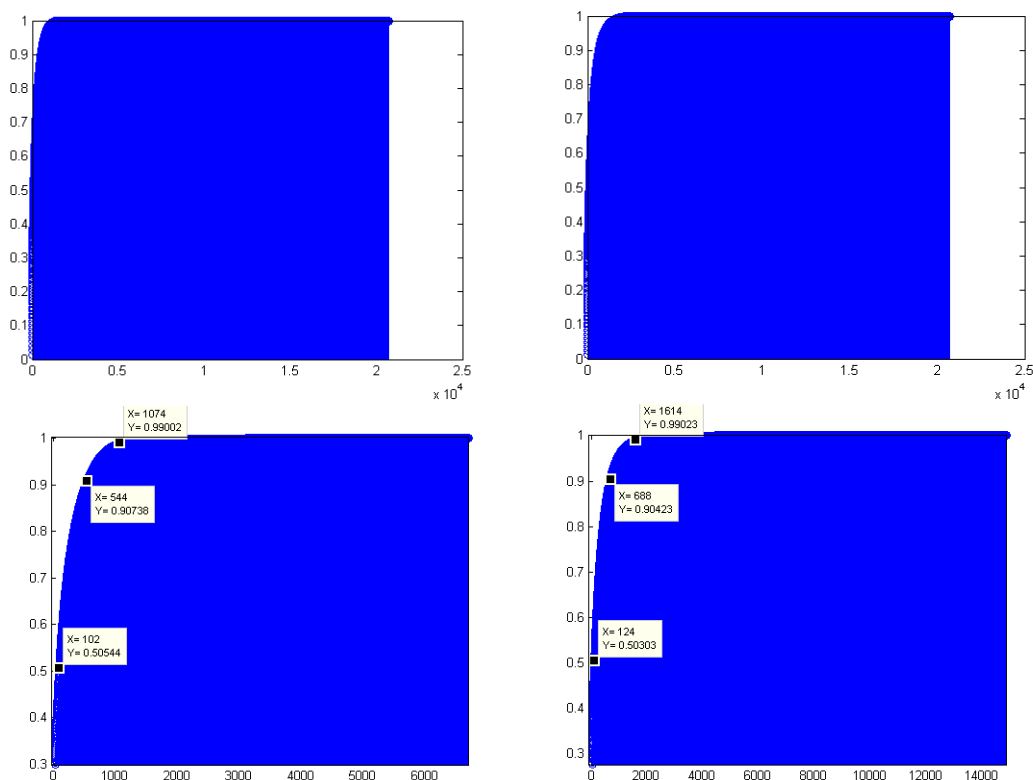
Com s'ha comentat anteriorment, per reflectir la incidència del valor d' $\varepsilon$ , es va realitzar el mateix estudi que s'ha mostrat amb els gràfics percentuals de l'energia respecte els coeficients amb un valor d' $\varepsilon = 10^{-15}$ .





**FIG. 5.9.** Coeficients amb  $\epsilon = 10^{-15}$  (esquerra), i  $\epsilon = 10^{-8}$  (dreta).

La figura 5.9 ens mostra les diferències entre els coeficients adquirits amb el mateix mètode d'anàlisi, en aquest cas opció 4, amb un valor d' $\epsilon$  diferent. Els dos gràfics són pràcticament idèntics a excepció de punts de concentració major de coeficients en el cas dels coeficients amb l'  $\epsilon = 10^{-15}$ .



**FIG. 5.10.** Superior esquerra: Energia respecte els coeficients mètode d'estudi 4 amb  $\epsilon = 10^{-15}$ . Superior dreta: Energia respecte els coeficients mètode d'estudi 4 amb  $\epsilon = 10^{-8}$ . Inferior esquerra: Ampliació amb valors energia amb  $\epsilon = 10^{-15}$ . Inferior dreta: Ampliació amb valors energia amb  $\epsilon = 10^{-8}$ .

La figura 4.10 mostra la diferent distribució de l'energia respecte el coeficients amb diferents valors d' $\epsilon$ .

% Energia	$\epsilon = 10^{-8}$ Abilene	$\epsilon = 10^{-15}$ Abilene
50%	0,602 %	0,491 %
90%	3,255 %	2,623 %
99%	7,764 %	5,179 %

**Taula 5.2.** Comparació percentual de quantitat de coeficients respecte l'energia amb diferents valors d'  $\epsilon$ .

Com ens mostra la taula 5.2 amb un valor d'  $\epsilon$  menor s'obtenen millors resultats amb el mateix mètode d'estudi. Es a dir, amb el valor més baix d'  $\epsilon$  l'energia amb menor quantitat de valors es capta més energia de la matriu original, i per tant menys valors calen per realitzar la reconstrucció d'aquesta

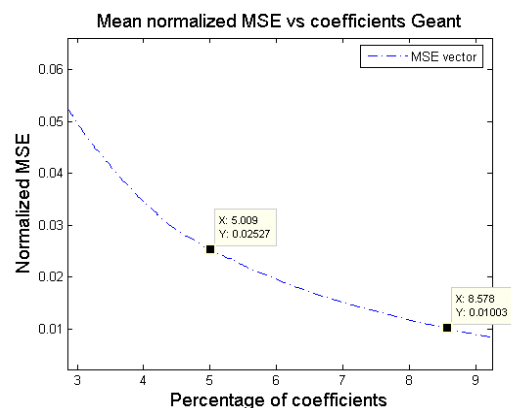
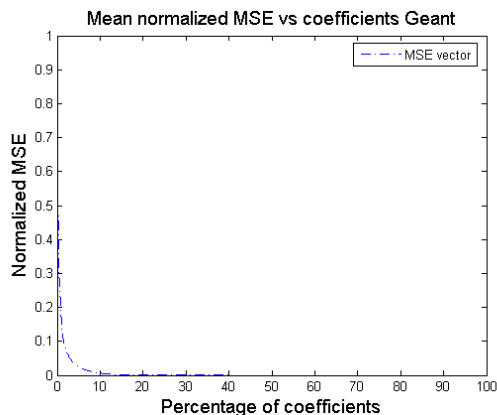
Tot i les dades aportades en aquesta taula, estudis anteriors amb diferents valors d'  $\epsilon$  demostren que aquesta relació no és lineal. El fet que el valor de la  $\epsilon$  disminueixi no assegura que la relació dels coeficients amb l'energia de la matriu millori. Aquest fet es farà evident amb l'estudi realitzat sobre Geant.

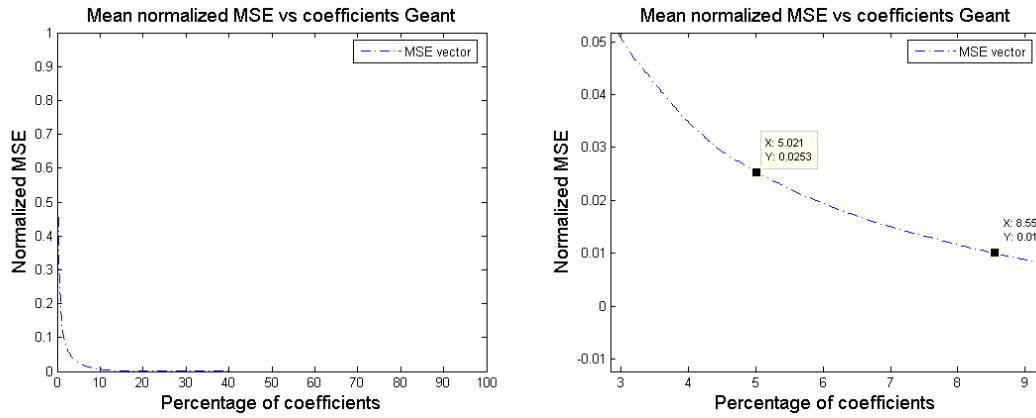
## 5.2. Geant

Respecte els resultats obtinguts amb Geant només s'ha avaluat amb l'opció 4, la que oferia millors resultats en el cas d'Abilene.

### 5.2.1. MSE

Com en el cas d'Abilene els primers resultats que es mostren són referents a l'error quadràtic mig.

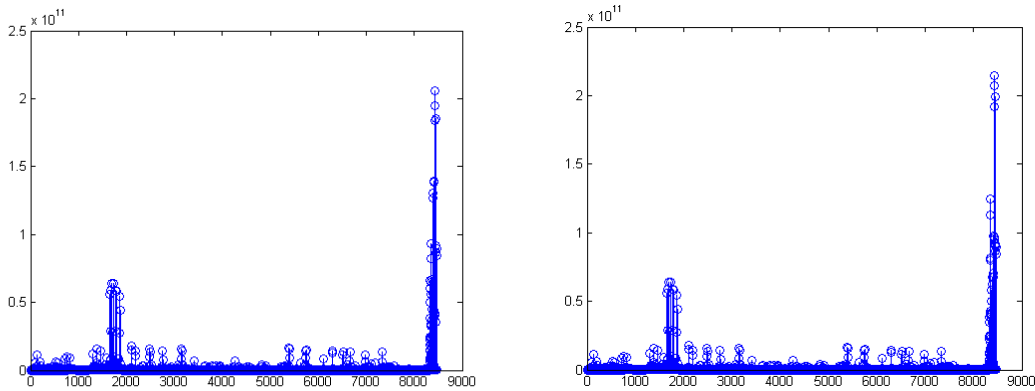




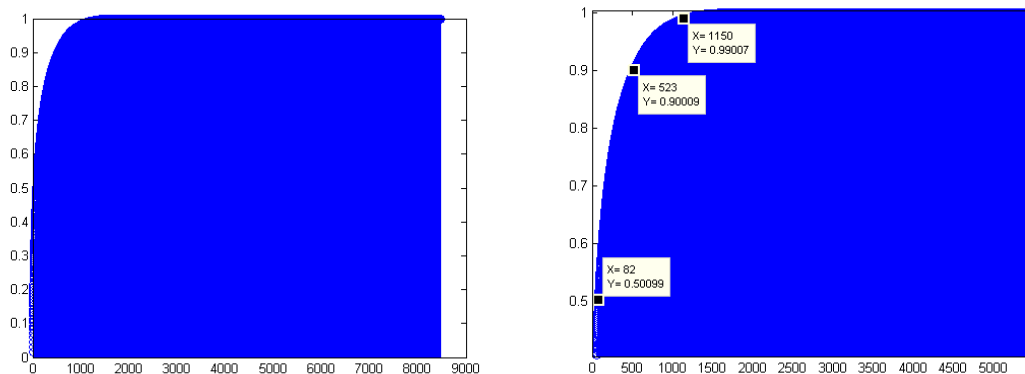
**FIG. 5.11.** Error quadràtic mig normalitzat en funció del percentatge de coeficients de la DW a Geant. Superior esquerra: MSE amb  $\epsilon=10^{-8}$ . Superior Dreta: Ampliació MSE amb  $\epsilon=10^{-8}$ . Inferior Esquerra: MSE amb  $\epsilon=10^{-15}$ . Inferior Dreta: Ampliació MSE amb  $\epsilon=10^{-15}$ .

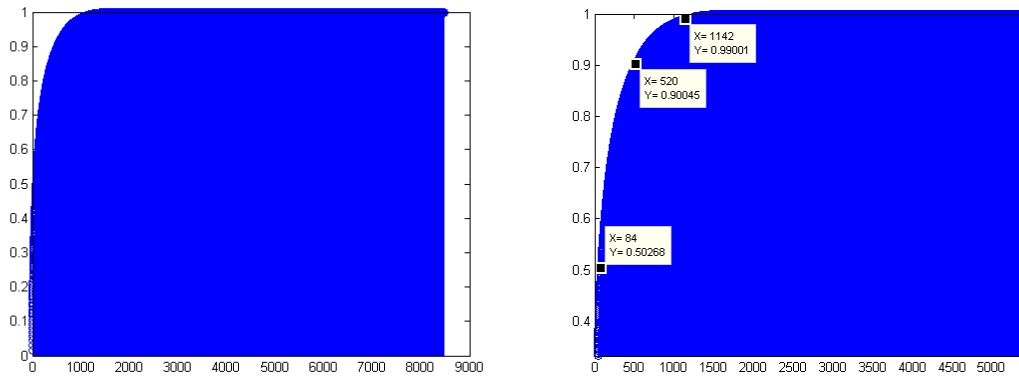
**5.2.2. Energia acumulativa segons coeficients**

En les figures 5.13 i 5.14 es pot apreciar la quantitat de coeficients resultants i el percentatge de l'energia respecte aquest coeficients.



**FIG. 5.13.** Coeficients mètode d'estudi 4 amb Geant amb valors d'  $\epsilon$  diferents Esquerra:  $\epsilon= 10^{-8}$ . Dreta:  $\epsilon= 10^{-15}$





**FIG. 5.14.** Superior esquerra: Energia respecte els coeficients  $\epsilon = 10^{-8}$ . Superior dreta: Ampliació gràfic energia respecte els coeficients  $\epsilon = 10^{-8}$ . Inferior esquerra: Energia respecte els coeficients amb  $\epsilon = 10^{-15}$ . Inferior dreta: Ampliació gràfic energia respecte els coeficients  $\epsilon = 10^{-15}$ .

A la taula 5.3 es realitzar una comparativa a partir dels gràfics anteriors, on es pot apreciar que, a diferència amb Abilene, el valor  $\epsilon$  no té la mateixa rellevància.

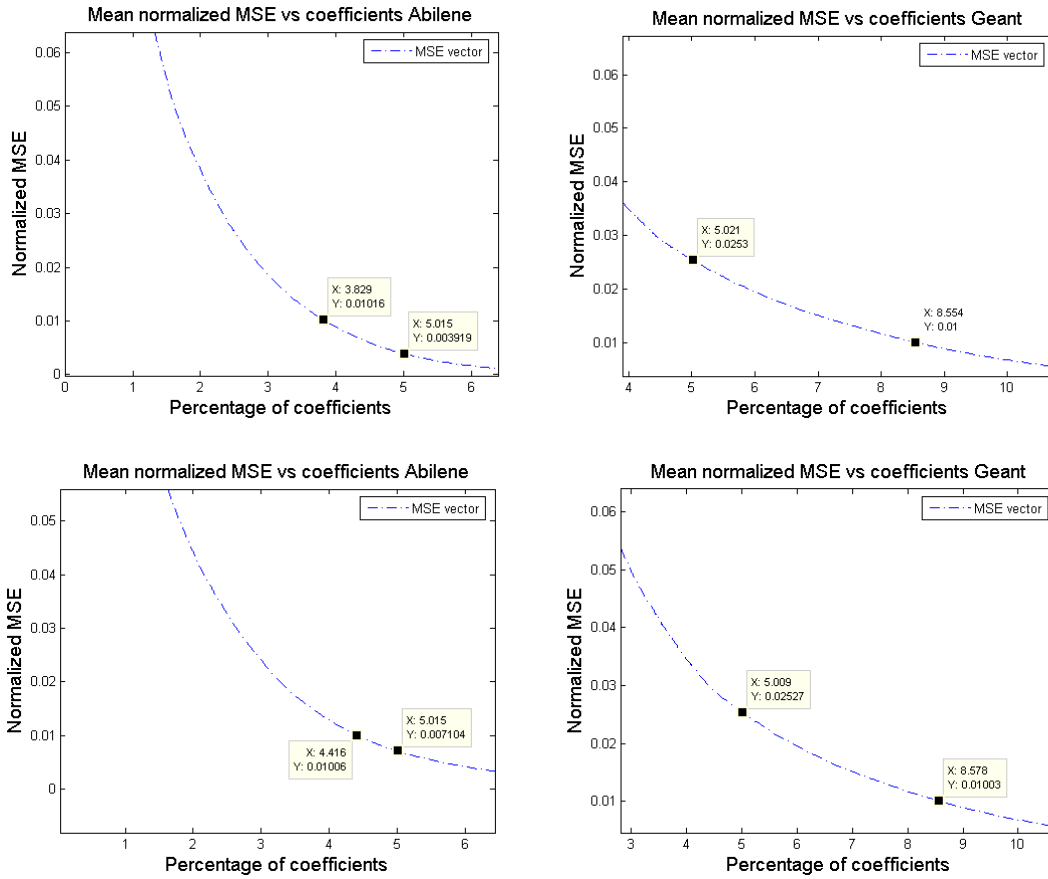
% Energia	$\epsilon = 10^{-8}$ Geant	$\epsilon = 10^{-15}$ Geant
50%	0,968 %	0,992 %
90%	6,179 %	6,131 %
99%	13,586 %	13,539 %

**Taula 5.3.** Comparació percentual de quantitat de coeficients respecte l'energia amb diferents valors d'  $\epsilon$ .

### 5.3. Comparativa Abilene vs Geant

#### 5.3.1. Comparativa MSE

La primera comparativa que es mostrar és respecte l'error quadràtic mig. En aquesta comparativa es pot apreciar que els resultats d'Abilene són sensiblement millors que els de Geant. En el cas d'Abilene com s'ha comentat anteriorment el valor de la  $\epsilon$  té una rellevància major que en el cas de Geant.

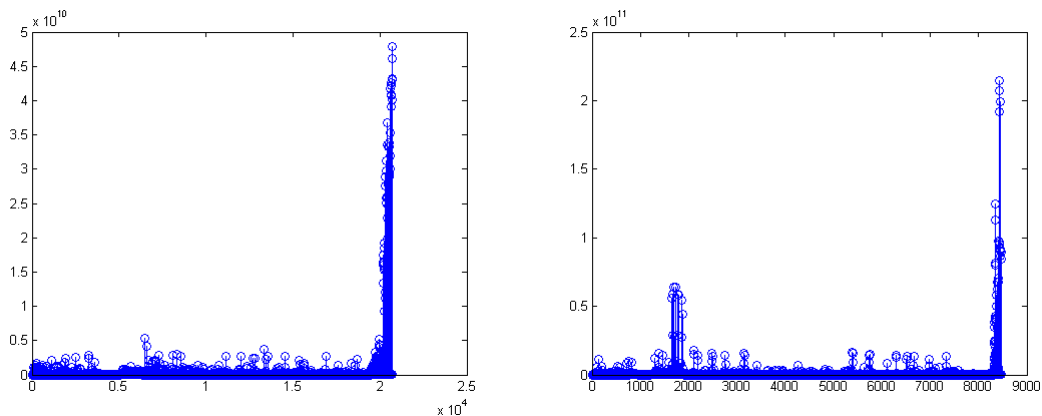


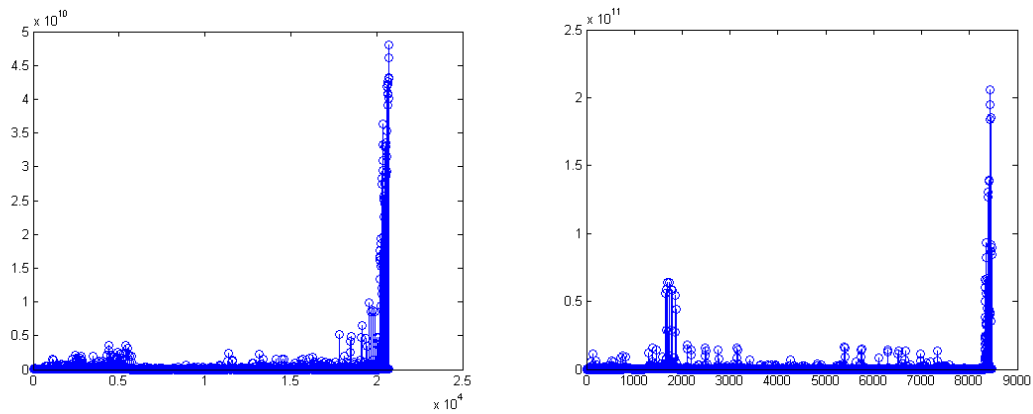
**FIG. 5.15.** Superior: Comparativa Abilene-Geant MSE amb  $\epsilon = 10^{-15}$ . Inferior: Comparativa Abilene-Geant MSE amb  $\epsilon = 10^{-8}$ .

La comparativa dels MSE's de cada una de les xarxes demostra que en el cas d'Abilene la incidència del valor d'  $\epsilon$  és més gran. Les diferències en els percentatges en el cas de Geant són imperceptibles.

### 5.3.2. Comparativa Coeficients

En aquesta comparativa, una vegada més s'aprecia que el valor de la  $\epsilon$  pren més rellevància en el cas d'Abilene. Els gràfics de Geant són pràcticament idèntics, mentre que en el cas d'Abilene, tot i ser molt semblants, es poden apreciar majors diferències.



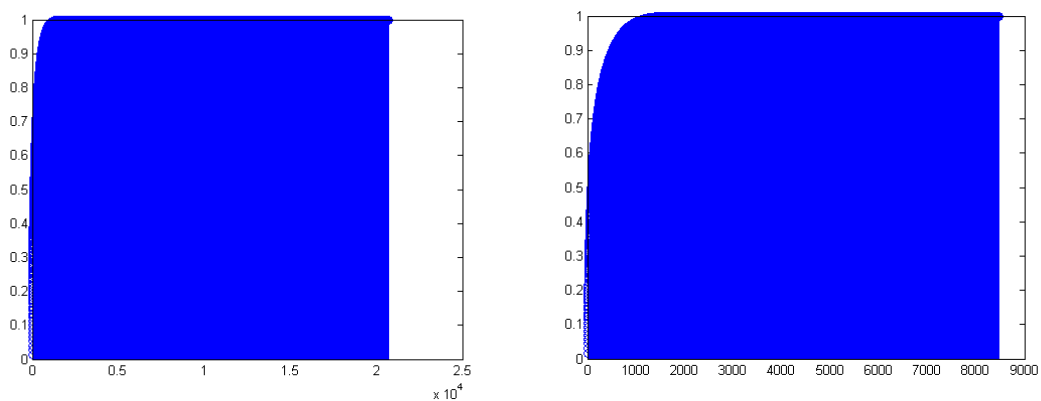


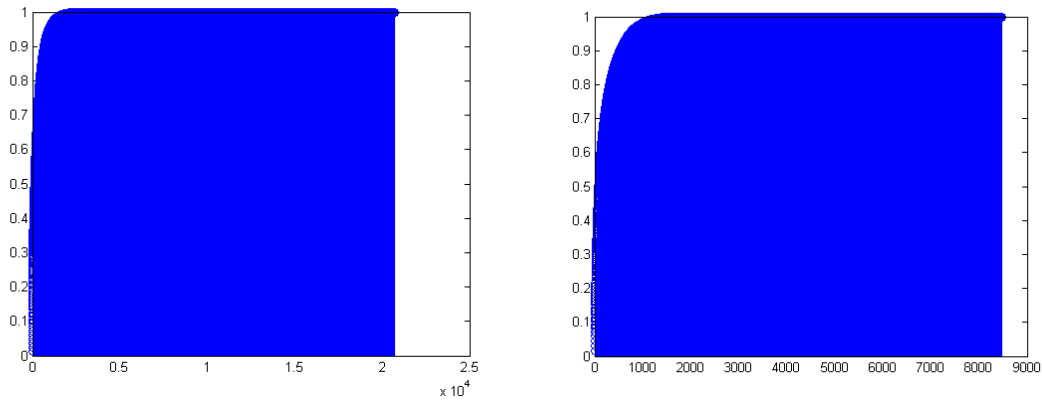
**FIG. 5.15.** Superior esquerra: Coeficients Abilene amb  $\varepsilon = 10^{-15}$ . Superior dreta: Coeficients Geant amb  $\varepsilon = 10^{-15}$ . Inferior esquerra: Coeficients Abilene amb  $\varepsilon = 10^{-8}$ . Inferior dreta: Coeficients Geant amb  $\varepsilon = 10^{-8}$ .

La comparativa amb els coeficients delata que en el cas de la xarxa Abilene la diferència dels valors d'  $\varepsilon$  ha sigut molt més rellevant. En el gràfics dels coeficients d'ambdues xarxes s'aprecia que en el cas de Geant, la distribució dels coeficients és pràcticament idèntica. Abilene, per la seva banda, mostra una certa densitat superior en alguns trams del gràfic amb una  $\varepsilon$  més baixa.

### 5.3.3. Comparativa energia respecte els coeficients

La comparativa de l'energia respecte els coeficients mostren clarament el que s'havia demostrat amb els anteriors gràfics comparatius. Els resultats són millors, en el cas d'Abilene quant  $\varepsilon$  té un valor menor. En el cas de Geant, no s'aprecien diferències.





**FIG. 5.16.** Superior esquerra: Energia respecte els coeficients Abilene amb  $\epsilon=10^{-15}$ . Superior dreta: Energia respecte els coeficients Geant amb  $\epsilon=10^{-15}$ . Inferior esquerra: Energia respecte els coeficients Abilene amb  $\epsilon=10^{-8}$ . Inferior dreta: Energia respecte els coeficients Geant amb  $\epsilon=10^{-8}$ .

Una vegada més els gràfics de la distribució de l'energia en els coeficients demostra clarament que en el cas de Geant no hi ha percepció de canvi. En el cas d'Abilene és molt més evident l'evolució dels resultats de l'energia. En aquest cas las taules 5.2 i 5.3 denoten una millora significativa en el cas d'Abilene, i cap evolució en el cas de Geant.

#### 5.3.4. Estudi de la mateixa hora en set dies consecutius

Donat els resultats mostrats fins el moment es va decidir realitzar un estudi sobre un espectre més ampli de matrius de teletrànsit. Per realitzar aquest estudi es va utilitzar únicament el mètode quatre de càlcul amb una  $\epsilon=10^{-15}$ , degut a que amb aquest valor d' $\epsilon$ , en el cas d'Abilene, s'havien assolit millors resultats. (Annexos, capítol 4).

Per realitzar aquest estudi es va procedir a avaluar l'evolució del teletrànsit durant una setmana, a la franja horària de les 8 a les 14 hores. Una vegada determinat que el teletrànsit no presentava anomalies es va determinar, en el cas d'Abilene, agafar com a hora d'estudi la franja de les 09:00 a 09:55. En el cas de Geant no es pot afirmar amb seguretat que las dades corresponguin a les hores indicades pels *-datasets-*, però es va realitzar la mateixa avaluació del teletrànsit que en el cas de la xarxa americana, i es va valorar utilitzar la franja horària de les 08:00 a 08:45 indicada pels *-datasets-*. Un primer fet en l'anàlisi del teletrànsit de les dues xarxes va ser l'evident diferenciació del comportament del teletrànsit en les dues xarxes, però com s'ha explicat amb anterioritat, no es poden extreure conclusions d'aquest fet degut a que les dades de Geant, amb un probabilitat elevada, no corresponen a la franja horària indicada pels *-datasets-*.

Una vegada determinada l' hora d'estudi a cada xarxa es va procedir a realitzar l'anàlisi vist als anteriors punts, amb els paràmetres indicats anteriorment, amb la utilització únicament del mètode quatre i amb un  $\epsilon=10^{-15}$ .

Els resultats des del punt de vista dels coeficients respecte l'energia són resumits a les taules 5.4 i 5.5 per a cadascuna de les xarxes.

<b>Abilene</b>	1 <sup>er</sup> Dia	2 <sup>on</sup> Dia	3 <sup>er</sup> Dia	4 <sup>rt</sup> Dia	5 <sup>e</sup> Dia	6 <sup>e</sup> Dia	7 <sup>e</sup> Dia
50 %	0,4292	0,4677	0,5401	0,2507	0,5256	0,5353	0,5304
90 %	2,3823	2,5366	2,58	2,1363	2,6475	2,6427	2,7488
99 %	5,0829	5,2951	5,2228	4,6923	5,3337	5,4108	5,488

**Taula 5.4.** Percentatge de quantitat de coeficients respecte l'energia durant la setmana del 05/07/2004 a 11/07/2004 a la franja horària de 09:00 a 09:55.

<b>Geant</b>	1 <sup>er</sup> Dia	2 <sup>on</sup> Dia	3 <sup>er</sup> Dia	4 <sup>rt</sup> Dia	5 <sup>e</sup> Dia	6 <sup>e</sup> Dia	7 <sup>e</sup> Dia
50 %	1,0633	1,0396	1,0160	1,1932	1,0633	1,0751	1,004
90 %	6,1909	6,0255	6,2854	6,5689	6,3917	5,7065	5,5765
99 %	13,7169	13,6814	13,9886	14,1895	13,5160	13,1498	12,3818

**Taula 5.5.** Percentatge de quantitat de coeficients respecte l'energia durant la setmana del 10/01/2005 a 16/01/2005, segons *-datasets-* de Geant, a la franja horària de 08:00 a 08:55.

Els resultats que es mostren a las taules 5.4 i 5.5 són pràcticament idèntics als referents a l'estudi realitzant amb anterioritat en un hora aïllada.



## CONCLUSIONS

En aquest document s'ha mostrat la importància de les matrius de teletrànsit en l'àmbit de la enginyeria de trànsit, la seva rellevància en la gestió i configuració d'aquestes xarxes.

Les aplicacions de les matrius de teletrànsit es poden trobar des de la configuració del balanceig de càrrega generat pels nodes de la xarxa, a la configuració de l'encaminament de la xarxa, la detecció de possibles anomalies i predicció d'aquestes.

Tot i la utilitat evident de les matrius de teletrànsit existeixen problemes a l'hora d'aconseguir matrius de xarxes reals. Per una banda està la dificultat d'obtenir dades en temps real sense afectar la capacitat computacional dels nodes. El mètode que aporta un compromís acceptable és la captura de les dades amb el protocol SNMP.

Un segon problema és degut a las reticències dels ISP comercials a publicar les dades de les seves xarxes, degut a la competència entre empreses. Aquest problema és el que genera un lentitud extraordinària a l'hora de realitzar investigació en el camp de les matrius de teletrànsit. En aquest estudi s'ha treballat amb dades de las xarxes americana, Abilene, i europea, Geant.

El model d'anàlisi utilitzat en aquest estudi s'ha basat en la multiresolució de les matrius de teletrànsit. S'ha realitzat una introducció a las diferents eines utilitzades per realitzar la multiresolució. Aquestes eines es basant amb les Wavelets de difusió. Aquest tipus de wavelets ens permet estendre l'anàlisi multiresolució aplicat en altres camps, com per exemple les imatges, al camp dels grafs. Amb les wavelets de difusió i un operador de difusió es pot realitzar l'anàlisi multiresolució. En el nostre cas aquest operador s'ha basat en el model de gravetat.

L'anàlisi multiresolució permet, una vegada descomposta la matriu de teletrànsit, pogué reconstruir-la. En el cas que aquests valors acumulin una quantitat d'energia significativa de la matriu original, permet treballar amb una quantitat de valors molt inferior als que conté una matriu de teletrànsit original.

L'objectiu ha estat millorar els resultats obtinguts en altres estudis, en quant a la compressibilitat de les TMs. La contribució que s'ha realitzat és la introducció d'un coeficient de correlació a l'estudi de les matrius de teletrànsit en un àmbit temporal. S'ha demostrat que l'aplicació del coeficient de correlació sobre l'operador de difusió ha millorat els resultats respecte estudis anteriors sobre altres operador de difusió.

El mode d'aplicació del coeficient de correlació, com s'ha demostrat amb l'estudi realitzat sobre la xarxa Abilene, no és trivial. Durant l'estudi s'han realitzat diferents introduccions del coeficient, mostrant que els resultats milloren o empitjorant segons la seva aplicació.

Tot i que els resultats, sobretot en el cas d'Abilene, són una millora respecte altres estudis, no es pot considerar com a definitius. La falta de capacitat computacional només ha permès resultats sobre un rang molt petit de dades. En el cas de les dues xarxes s'ha realitzat un estudi màxim d'una hora durant set dies consecutius, per tant es pot considerar com insuficient per extreure conclusions més enllà de la evident millora, com s'ha comentat, en el cas de la xarxa americana.

El futur dels estudis hauran de partir de la idea d'obtenir resultats amb un espectre molt més gran del que s'ha pogut realitzar en aquest estudi. Seria important obtenir resultats d'hores consecutives. Aquest fet podrien confirmar els bons resultats obtinguts en aquest estudi.

A la referència [35] es van mostrar els primers resultats de l'estudi realitzat en aquest treball, referents a la primera hora sobre la que es va fer l'anàlisi. Com s'explica a l'article, i posteriorment s'ha pogut confirmar amb més hores estudiades, els resultats són prometedors, tot i que faran falta estudis més amplis per concretar la millora.

Des de el punt de vista de l'ambientalització s'ha de tenir en compte que el fet de tenir millors mètodes d'encaminament del teletrànsit permet estalviar energia en els "routers" i altres elements de la xarxa. El fet d'evitar col·lapses a les xarxes genera que no s'hagi de reenviar dades ni fer sobreesforços per part dels elements de la xarxa per redirigir el teletrànsit. Tant el reenviament de dades com els càlculs d'encaminament suposen un consum energètic. Amb estudis com el presentat en aquest treball l'estalvi energètic generat a les xarxes pot arribar a ser apreciable.

## BIBLIOGRAFIA

- [1] Roughan, M. Thorup, M. and Zhang, Y., "Traffic Engineering with Estimated Traffic Matrices", *Proceedings of the USENIX/ACM Internet Measurement Conference (Miami, Florida)*. pp. 248-258. (2003).
- [2] Leprope, J., Balon, S. and Leduc, G., "TOTEM: a TOolbox for Traffic Engineering Methods", Poster and Demo Session of *INFOCOM'06 (Barcelona, Spain)*. (2006).
- [3] Balon, S., Leprope, J., Delcourt, O., Skivée, F. and Leduc, G., "Traffic Engineering an Operational Network with the TOTEM Toolbox", *IEEE Transactions on Network and Service Management*. 4 (1). pp. 51-61. (2007).
- [4] Coates, M., Pointurier, Y. and Rabbat, M., "Compressed network monitoring for IP and all-optical networks", *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement (San Diego, California)*. pp. 241-252. (2007).
- [5] Crovella, M. and Kolaczyk, E., "Graph Wavelets for Spatial Traffic Analysis", *INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications. IEEE Societies*. 3. pp. 1848-1857. (2003).
- [6] Rincón, D., Roughan, M., and Willinger, W., "Towards a Meaningful MRA of Traffic Matrices", *Internet Measurement conference. Proceedings of the 8th ACM SIGCOMM conference on Internet Measurements (Vouliagmeni, Greece)*. pp. 331-336. (2008).
- [7] Wild, M., "Nonlinear Approximation of Spatiotemporal Data Using Diffusion Wavelets", in *Computer Analysis of Images and patterns*. Springer Berlin/Heidelberg. 4673. pp. 886-894. (2007).
- [8] García de Jalón, J., Rodríguez, J.I. and Vidal, J., "Aprenda Matlab 7.0 como si estuviera en primero", *Tutoriales UPM*. (2006).  
<http://meteo.fisica.edu.uy/?download=matlab70primero.pdf>
- [9] Coifman, R.R. and Maggioni, "Diffusion Wavelets", in *Applied and Computational Harmonic Analysis*. 21 (1). pp. 53-94. (2006).
- [10] Rincón, D., Torres, J., "Contribuciones al análisis multirresolución de matrices de tráfico", *VIII Jornadas de Ingeniería Telemática JITEL 2009 (Cartagena, Spain)*. pp. 413-420. (2009).
- [11] Lakhina, A., Papagiannaki, K., Crovella, M., Diot, C., Kolaczyk, E. and Taft, N., "Structural Analysis of Network traffic Flows", *Joint International Conference on Measurement and Modeling of Computer Systems (New York, USA)*. pp. 61-76. (2004).

[12] Maggioni, M., "Diffusion analysis of and on graphs and high dimensional data", Invited talk at *IPAM-UCLA*, (2007).

<http://www.math.duke.edu/~mauro/Talks/Banff.pdf>

[13] Zhang, Y., Roughan, M., Duffield, N., Greenberg, A, "Fast Accurate Computation of Large Scale IP Traffic Matrices from Link Loads", *ACM SIGMETRICS Performance Evaluation Review (New York, USA)*. 31 (1). pp. 206-217. (2003).

[14] Ceuppens, L., Sardella, A., Kharitonov, D., "Power Saving Strategies and Technologies in Network Equipment Opportunities and Challenges, Risk and Rewards", *Proceedings of the 2008 International Symposium on Applications and the Internet*. pp. 381-384. (2008).

[15] Uhlig, S., Quoitin, B., Lepropre, J., Balon, S., "Providing public intradomain traffic matrices to the research community", *ACM SIGCOMM Computer Communication review*. 36 (1). pp 83-86. (2006).

[16] Mallat, S., *A wavelet tour of signal processing*. Academic Press, London 1999.

[17] Fortz, B., Rexford, J., Thorup, M., "Traffic engineering with traditional IP routing protocols", *IEEE Communication Magazine*. 40 (10). pp. 118-124. (2002)

[18] Nucci, A., Sridharan, A., Taft, N., "The problem of synthetically generating IP traffic matrices: Initial recommendations", *SIGCOMM Computer Communication Review*. 35 (3). (2005).

[19] F. Chung. "Spectral Graph Theory", *CBMS Regional Conference Series in Mathematics*. 92. (1997).

[20] Universidad Nacional del Centro de Buenos Aires (UNICEN, Argentina), "Introducción a la transformada Wavelet"

<http://www.exa.unicen.edu.ar/escuelapav/cursos/wavelets/apunte.pdf>

[21] Mathwork's Matlab Home Site <http://www.mathworks.com/products/matlab>

[22] TOTEM Project Home Site <http://totem.run.montefiore.ulg.ac.be>

[23] Roughan, M., "AT&T Labs Tomogravity research"

<http://research.att.com/viewProject.cfm?projID=133>

[24] Wikipedia, *Spectral Theorem* [http://en.wikipedia.org/wiki/Spectral\\_theorem](http://en.wikipedia.org/wiki/Spectral_theorem)

[25] Wikipedia, *Discrete Wavelet Transform*

[http://en.wikipedia.org/wiki/Discrete\\_wavelet\\_transform](http://en.wikipedia.org/wiki/Discrete_wavelet_transform)

[26] Wikipedia, *Wavelet Packet Decomposition*

[http://en.wikipedia.org/wiki/Wavelet\\_packet\\_decomposition](http://en.wikipedia.org/wiki/Wavelet_packet_decomposition)

- [27] Simple Network Management Protocol (SNMP) RFC  
<http://tools.ietf.org/html/rfc1157>
- [28] Wikipedia, *NetFlow* <http://en.wikipedia.org/wiki/Netflow>
- [29] Wikipedia, *Multiresolution Analysis*  
[http://en.wikipedia.org/wiki/Multiresolution\\_analysis](http://en.wikipedia.org/wiki/Multiresolution_analysis)
- [30] JPEG 2000 RFC <http://tools.ietf.org/html/rfc3745#ref-ISO-JPEG2000-1>
- [31] Wikipedia, *Point of Presence* [http://en.wikipedia.org/wiki/Point\\_of\\_presence](http://en.wikipedia.org/wiki/Point_of_presence)
- [32] Abilene Network <http://www.internet2.edu/network/>
- [33] GÉANT Network <http://www.geant.net/>
- [34] Josep Xavier Torres Haba, “*Modelling of traffic matrices with multiresolution analysis techniques*”. Treball Fi de Carrera EPSC (Escola politècnica Superior de Castelldefels)(2009)
- [35] Rincón, D.,Balasch, I., “Análisis espacio-temporal multiresolución de matrices de tráfico”. Enviat a les IX Jornadas de Ingenieria Telemática Jitel 2010 (Valladolid) (2010)

## GLOSSARI

**2D DW:** Wavelets de difusió en dos dimensions, (*"Diffusion wavelets in two dimensions"*).

**DoS:** Denegació de Servei, (*"Deny of Service"*).

**DW:** Wavelet de difusió, (*"Diffusion Wavelet"*).

**DWT:** Transformada Wavelet Discreta, (*"Discrete Wavelet Transform"*).

**GM:** Model de Gravetat, (*"Gravity Model"*).

**ISP:** Proveïdor del servei d'Internet Internet Service Provider, (*"Intrenet Service Provider"*).

**MAN/WAN:** Xarxa d'Àrea Metropolitana / Ampla, (*"Metropolitan / Wide Area Network"*).

**MRA:** Anàlisi Multiresolució, (*"MultiResolution Analysis"*).

**MSE:** Error Quadràtic Mig, (*"Mean Square Error"*).

**PoP:** "Point of Presence"

**QoS:** Qualitat de Servei, (*"Quality of Service"*).

**SNMP:** Protocol de Gestió de Xarxa Simple, (*"Simple Network Management Protocol"*).

**TM:** Matriu de Teletrànsit, (*"Traffic Matrix"*).

**TOTEM:** Toolbox for Traffic Engineering Methods