Universitat Autònoma de Barcelona

Departament de Geografia

Facultat de Filosofia i Lletres

Tesi doctoral

Doctorat en Geografia

Contribucions a l'anàlisi automàtica de grans volums de dades en el context de la Ciència de la Informació Geogràfica

Contributions to the automatic analysis of big data in the context of GI Science

Doctorand:
Lluís Pesquer Mayos

Director:
Xavier Pons Fernández

2014

...a la Rosa

Índex

Agraïments

Són moltes les persones que mereixen que agreixi el seu suport i col·laboració en aquesta tesi, sóc molt afortunat d'haver tingut moltes companyes i companys ajudant-me.

Cronològicament haig de començar per agrair a les companyes i companys del DACSO de la UAB, tot han estat facilitats per usar les seves infraestructures informàtiques, essencials en diversos treballs d'aquesta tesi. I en aquest departament, mereix un agraïment força especial el suport incondicional de l'Anna Cortés. També de l'ETSE, però d'un altre departament, el GICI, vull donar les gràcies al Joan S.

També he rebut un suport molt gran del Departament de Geografia, on tinc els meus companys i amics del grup de recerca GRUMETS. Moltes gràcies, tinc mil motius per donar-vos les gràcies, Alaitz, Cristina D. (desitjo que molt aviat estiguis escrivint uns agraïments semblants), Cristina C., Oscar, Pere... i també és imprescindible afegir en aquest pack a altres grumets, el Jordi (la seva ajuda ha estat proporcional a la seva radiació), Miquel, Ricardo, etc.

I el CREAF? Tinc ara mateix una sensació agredolça, d'agraïment infinit a algunes persones, com el cas del Gerard (necessitaria un annex especial si volgués detallar els motius), i de tristesa perquè quan vaig començar a fer la tesi hi havia uns companys de treball que la crisi i els seus efectes secundaris han provocat que ja no puguin treballar-hi. Voldria donar les gràcies als companys del mapa de cobertes, a l'Anna R., Edu, Xavier C., Victor...i desitjar-los tota la sort del món. Però al CREAF continua quedant gent fantàstica, els companys del futbol, els *runners*, els *trailwalkers*, etc, que m'han ajudat a desconnectar de tant en tant, i com els grumets més propers: Arnald, Núria, Abel, Ester, Ivette, Joan M. (voltant pel món, però molt a prop quan fa falta). Al CREAF oficial també m'ha donat suport i recursos, gràcies Javi i gràcies Joan P. (on tot va començar?... gràcies Pino).

Els agraïments al Xavier P. són transversals com aquesta tesi. El Xavier ha deixat que jo portés la tesi per on desitjava, però sempre que ho he necessitat m'ha aconsellat i dirigit; ha tret hores d'on podia per a revisar a fons els treballs que formen la tesi i ha posat recursos quan calia. I en els moments complicats (crec que en una tesi tothom en té) m'ha donat els ànims que necessitava, moltes gràcies Xavier!

I la família? Escriure els agraïments a la família són durs, perquè res no pot tornar el temps robat. He intentat minimitzar els efectes col·laterals de la tesi, però..¿?... no són tant la càrrega de feina addicional que ha "tocat" a casa, com alguna època de cansament, desànim meu, etc. que alguna vegada els pot haver afectat. MOLTES GRÀCIES per la comprensió Maribel, Martí i Laia.

Ah! ...i la portada ha estat gràcies a una idea de la Roser, que el José Luis ha acabat de desenvolupar. Gràcies!!!

Resum

Es presenta una tesi multidisciplinària en el context de la Ciència de la Informació Geogràfica on convergeixen sinèrgicament metodologies de la Geostadística, la Teledetecció, la Ciència Computacional i els Sistemes d'Informació Geogràfica amb l'objectiu d'estudiar i oferir noves solucions que permetin l'automatització de determinats algorismes d'anàlisi espacial i de processament d'imatges. L'automatització d'aquests algorismes està enfocada a possibilitar i millorar l'eficiència del processament massiu de grans volums de dades geospacials.

L'automatització d'aquests processos està fonamentada en dos pilars principals:

- L'aprofitament adequat de les metadades en la presa de decisions automàtiques. Aquestes metadades poden ser generades tant pel proveïdor de les dades originals com pels resultats intermedis derivats dels processos implicats en els algorismes estudiats. Aquest important rol de conductor dels processos d'anàlisi espacial afegeix a les metadades un component més geogràfic que altres usos més habituals com són les cerques sobre catàlegs.
- L'estudi i adaptació específic de l'algorisme a l'automatització. S'ha analitzat i re-dissenyat cada algorisme fins aconseguir una solució automàtica que mantingui la màxima qualitat possible: *kriging*, anàlisis del variograma, interpoladors òptims, models de regressió multivariant o logística adequats i correccions radiomètriques.

Addicionalment, l'automatització ha permès aplicar les metodologies de distribució de còmput entre diferents unitats de processament i reduir de forma significativa els temps d'execució (s'ha aconseguit *speed-ups* propers a 60 en algun cas, i propers a un comportament lineal en altres). Aquestes millores computacionals s'han dut a terme mitjançant la paral·lelització dels codis en llenguatge C i l'enviament de dades entre processadors mitjançant Message Passing Inteface (MPI) en un entorn d'elevades prestacions computacionals (HPC).

Les automatitzacions s'han aplicat a diversos camps i a diversos tipus de dades, essent les sèries temporals d'imatges de Teledetecció el principal àmbit de recerca i aplicació. Per exemple, en aquesta tesi es proposa un mètode que millora la selecció d'imatges (reflectàncies de Terra MODIS) d'elevada qualitat respecte a la que només s'efectua en funció de les màscares de qualitat que acompanyen aquests productes. El nou mètode afegeix una anàlisi automàtica del variograma que detecta les possibles anomalies radiomètriques i geomètriques respecte un patró espacial tipus, definit també amb criteris geostadístics.

Aconseguir corregir radiomètricament de forma automàtica imatges Landsat de diversos sensors en base a valors radiomètrics de referència des de polígons pseudo-invariants, també generats automàticament, és una altra de les fites d'aquesta tesi. Aquesta aportació permet processar llargues sèries temporals d'imatges d'elevada resolució espacial per a àmbits geogràfics extensos amb un elevat rigor radiomètric i, per tant, obre la porta a realitzar anàlisis del territori amb un nivells de detall fins ara no viables.

Finalment, en el camp de la compressió amb pèrdua d'imatges de Teledetecció s'ha realitzat aportacions que afegeixen una visió més espacial (i per tant territorial) sobre visions més clàssiques en l'àmbit general del tractament d'imatges. La compressió d'imatges és una de les metodologies àmpliament acceptades avui per a gestionar de forma eficient grans volums de dades i, altre cop, l'anàlisi geostadística ha permès conduir les compressions cap a resultats que conservin millor les propietats espacials de les imatges comprimides.

Resumen

Se presenta una tesis multidisciplinar en el contexto de la Ciencia de la Información Geográfica dónde convergen sinérgicamente metodologías de la Geoestadística, la Teledetección, la Ciencia Computacional y los Sistemas de Información Geográfica con el objetivo de estudiar y ofrecer nuevas soluciones que permitan la automatización de determinados algoritmos de análisis espacial y de procesamiento de imágenes. La automatización de estos algoritmos está enfocada a posibilitar y mejorar la eficiencia del procesamiento masivo de grandes volúmenes de datos geoespaciales.

La automatización de estos procesos está fundamentada en dos pilares principales:

- El aprovechamiento adecuado de los metadatos en la toma de decisiones automáticas. Estos metadatos pueden ser generados tanto por el proveedor de los datos originales como por los resultados intermedios derivados de los procesos implicados en los algoritmos estudiados. Este importante rol de conductor de los procesos de análisis espacial añade a los metadatos una componente más geográfica que otros usos más habituales como son las buscas sobre catálogos.
- El estudio y adaptación específico del algoritmo a la automatización. Se ha analizado y rediseñado cada algoritmo hasta conseguir una solución automática que mantenga la máxima calidad posible: *kriging*, análisis del variograma, interpoladores óptimos, modelos de regresión multivariante o logística adecuados y correcciones radiométricas.

Adicionalmente, la automatización ha permitido aplicar las metodologías de distribución de cómputo entre diferentes unidades de procesamiento y reducir de forma significativa los tiempos de ejecución (se ha conseguido *speed-ups* próximos a 60 en algún caso, y próximos a un comportamiento lineal en otras). Estas mejoras computacionales se han llevado a cabo mediante la paralelización de los códigos en lenguaje C y el envío de datos entre procesadores mediante Message Passing Inteface (MPI) en un entorno de elevadas prestaciones computacionales (HPC).

Las automatizaciones se han aplicado a varios campos y a varios tipos de datos, siendo las series temporales de imágenes de Teledetección el principal ámbito de investigación y aplicación. Por ejemplo, en esta tesis se propone un método que mejora la selección de imágenes (reflectividades de Terra MODIS) de elevada calidad respecto a la que sólo se efectúa en función de las máscaras de calidad que acompañan estos

productos. El nuevo método añade un análisis automático del variograma que detecta las posibles anomalías radiométricas y geométricas respecto a un patrón espacial tipo, definido también con criterios geoestadísticos.

Conseguir corregir radiométricamente de forma automática imágenes Landsat de varios sensores en base a valores radiométricos de referencia desde polígonos pseudo-invariantes, también generados automáticamente, es otro de los logros de esta tesis. Esta aportación permite procesar largas series temporales de imágenes de elevada resolución espacial para ámbitos geográficos extensos con un elevado rigor radiométrico y, por lo tanto, abre la puerta a realizar análisis del territorio con un niveles de detalle hasta ahora no viables.

Finalmente, en el campo de la compresión con pérdida de imágenes de Teledetección se ha realizado aportaciones que añaden una visión más espacial (y por lo tanto territorial) sobre visiones más clásicas en el ámbito general del tratamiento de imágenes. La compresión de imágenes es una de las metodologías ampliamente aceptadas hoy para gestionar de forma eficiente grandes volúmenes de datos y, nuevamente, el análisis geoestadístico ha permitido conducir las compresiones hacia resultados que conserven mejor las propiedades espaciales de las imágenes comprimidas.

Abstract

This is a multidisciplinary thesis in Geographical Information Science. Methodologically, it draws on Geostatistics, Remote Sensing, Computer Science and Geographical Information Systems. The thesis studies several algorithms for spatial analysis and image processing, and proposes new ways to automate them. This automation is meant to enable and improve the efficiency of massive processing of big geospatial data.

The automation of these processes is based on:

1) The use of metadata in automated decision making. This metadata can be generated either by the supplier of the original data or by the intermediate results derived from the processes involved in the algorithms. Allowing them to drive the processes of spatial analysis gives metadata a geographic dimension that is missing in more common uses, such as catalogue search.
2) The study and specific adaptation of the algorithm to its automation. Each algorithm has been studied and redesigned to achieve an automatic solution while preserving the highest possible quality: kriging, variogram analysis, optimal interpolation, multivariant or logistic regression models, and radiometric corrections.

In addition, automation allows to distribute computing amongst different processing units, significantly reducing execution time (speed-ups close to 60 have been achieved in some instances, while in others behaviour was nearly linear). These computational improvements result from parallel programming in C language and from sending data between processors through Message Passing Interface (MPI) in a High Performance Computing (HPC) environment.

Automation has been applied to various fields and various types of data, first and foremost to time series of Remote Sensing images, our main domain of research and application. For example, this thesis proposes a method that improves the selection of high quality images (Terra MODIS reflectance products) with respect to the selection based on the quality masks accompanying these products. The new method adds an automatic variogram analysis capable of detecting possible radiometric and geometric anomalies with respect to a spatial pattern type, defined with geostatistical criteria.

Another aim of this thesis is to automate the radiometric correction of Landsat imagery from various sensors on the basis of radiometric reference values from pseudo-invariant polygons that are also automatically generated. This contribution makes it

possible to process long time series of high spatial resolution images covering large geographic areas, and to do so with high radiometric accuracy. This could be applied to geographical analysis over large areas at a level of detail not feasible until now.

Finally, in the field of lossy compression applied to Remote Sensing imagery, the thesis proposes a view that is more spatial (and therefore geographical) than the more traditional views in the domain of image processing. Image compression is widely accepted today as a methodology to efficiently manage big data. Here again, geostatistical analysis has allowed us to compress images with a minimum loss of their spatial properties.

1. Introducció

1.1 Introducció general

Els ordinadors van sorgir amb l'objectiu de poder realitzar càlculs que manualment era pràcticament impossible de fer, però també per a estalviar a les persones determinades tasques repetitives, permetre de fer càlculs mecanitzadament, etc, i d'aquesta forma, deixar a les persones la ment més lliure per a tasques més inventives, per a interpretar, analitzar, reflexionar, etc, i naturalment, per a efectuar la supervisió última de les tasques que realitzen els ordinadors.

Amb el pas del temps, aquesta situació s'ha repetit en algunes ocasions en un segon nivell. O sigui, els ordinadors realitzen tasques repetitives governades per un operador (humà d'entrada) que ordena les execucions, modifica els paràmetres, revisa els resultats, pren decisions derivades d'aquests resultats, etc. En canvi, una vegada consolidat el procés, el paper d'aquest operador pot anar perdent protagonisme, la supervisió pot tornar-se sistemàtica i la presa de decisions previsible. Podem donar una volta de cargol més i automatitzar aquestes tasques repetitives, preparant les decisions perquè coneixem el ventall de possibilitats dels resultats i les variables que els condicionen? Podem novament automatitzar les tasques automàtiques, tornant a deixar a la persona més disponible per a tasques on sigui més profitosa la seva expertesa?

El treball que presentem és una recerca en aquesta línia, en l'automatització, o segons el raonament anterior, en la ***metautomatització*** de determinats càlculs computacionals en l'àmbit de la Ciència de la Informació Geogràfica. L'objectiu és doncs, estudiar si podem delegar més rols a les màquines, perquè les persones tinguin temps per a analitzar, crear, imaginar, reflexionar, etc.

Tanmateix, els processos governats per les màquines tenen naturalment algunes restriccions importants, essent-ne la principal la manca de resposta amb una solució correcta a una situació no prevista (no tenen capacitat d'improvisació), però també tenen alguns avantatges: objectivitat en la presa de decisions (si ho són els criteris implicats), repetitivitat, traçabilitat (si ho permet un disseny i implementació apropiats en

les metadades), constància i homogeneïtat (la qualitat del procés, l'atenció en termes humans no decau per tasques repetitives), de manera, que si hem traslladat correctament el fenomen d'estudi a un model vàlid, podrem analitzar les diferents respostes en base a repeticions on canviïn uns pocs criteris o paràmetres, podrem replicar-los a altres àmbits o moments en el temps, analitzar diferents escenaris, establir causalitats, analitzar correlacions, dependències, etc. Ara bé, perquè un procés automàtic pugui ser alhora automatitzat per a diversos escenaris, (la metautomatització referida anteriorment) cal un important esforç en enriquiment de les metadades. Sense les metadades, no aconseguirem traçabilitat, ni tampoc la presa de decisions podrà respondre a criteris objectius ni aconseguirem situacions reproduïbles. Veurem, doncs, que una característica repetida, també d'una manera sistemàtica (encara que això no vulgui dir que el contingut temàtic d'aquesta tesi l'hagi ideat un ordinador, és clar) en els diferents capítols d'aquesta tesi, és el paper força rellevant de les **metadades**.

La Geografia en el seu conjunt, però especialment, les seves disciplines més quantitatives no han estat alienes als canvis de l'era digital comuns a altres àmbits de la Ciència, com ho són la informatització dels seus processos , algorismes, protocols i la digitalització de les seves fonts d'informació. En canvi, la recerca en l'automatització d'aquests processos és relativament incipient:

> *automated mapping is an emerging research field and will receive a significant attention in geography and Earth sciences in general* (Hiemstra *et al.* 2007).

i el treball que presentem té com una de les seves principals finalitats realitzar contribucions científiques en aquesta automatització.

També la Geografia, com molts altres àmbits, ha estat inundada per una enorme quantitat i diversitat de dades, el que s'anomena ***Big Data***. Cada vegada disposem (i també cada vegada de forma més assequible en termes econòmics) de cartografia de més detall, de zones més extenses i de temàtiques més diverses, de bases de dades més voluminoses tant pel nombre de mostres (registres) com per un gran nombre d'atributs i variables associades (camps), d'un enorme repositori d'ortofotografia aèria, d'un gran banc d'imatges de teledetecció, d'un important ventall de sensors, de models digitals del terreny de resolucions espacials de moltíssim detall, etc, i encara que els ordinadors van progressant en prestacions i potència de càlcul, ni de lluny s'acosten al ritme desbocat de l'increment de dades. Per tant, per una banda cal poder processar de forma eficient aquests grans volums de dades (i metadades), però també calen metodologies que sintetitzin, que ens donin la part més essencial de la informació que conté aquest gran volum de dades, estalviant-nos el processament d'informació redundant en alguns casos, o poc rellevant en altres casos.

La Ciència Computacional ens ofereix solucions per realitzar de forma correcta i eficient aquestes execucions de grans volums de dades i per reduir els seus temps d'execució mitjançant l'ús adequat d'entorns d'altes prestacions computacionals (HPC, per *High Performance Computing*). I, un cop més, l' automatització és una condició necessària per abordar el processament de grans volums de dades.

També en aquesta tesi s'estableix una aproximació per alleugerir el pes d'aquests grans volums de dades: els mètodes de compressió de dades amb pèrdua, en aquest cas aplicats a les imatges de Teledetecció, una de les importants fonts de *Big Data* en l'àmbit de la Ciència de la Informació Geogràfica.

Aquest treball no és una tesi purament de Teledetecció, però conté diferents capítols especialitzats en Teledetecció. Tampoc no té un enfocament estrictament de Ciència Computacional, però les seves metodologies esdevenen molt necessàries en diverses de les seves parts. El denominador comú d'aquesta tesi és el treball amb grans volums de dades en el context de la Ciència de la Informació Geogràfica i l'automatització rigorosa del seu tractament. Sí que podem considerar que és un treball amb una presència molt transversal de Geostadística (una disciplina present a tots els capítols), encara que les principals aportacions no són exactament pel fet d'aportar nous algorismes geostadístics, sinó per, o bé incorporar automatitzacions complexes en algorismes que necessiten una determinada presa de decisions clàssicament manual o interactiva, o per usar eines geostadístiques en l'anàlisi i processament de grans volums de dades. La sovint pesada maquinària geostadística ha estat majoritàriament associada a mostres poc nombroses en interpolacions i, en Teledetecció, a poques imatges o fins i tot, a fragments d'imatges de dimensions reduïdes (Chica-Olmo and Abarca-Hernandez 1998; Oliver *et al.* 2005; Rodgers and Oliver 2007). Com hem dit, podem contextualitzar perfectament el treball que presentem en l'àmbit de la Geografia, no només perquè és una tesi enterament dedicada a la Ciència de la Informació Geogràfica (CIG), un cop superada la seva antiga marginació respecte a la Geografia (Chuvieco *et al.* 2005), sinó també perquè l'anàlisi del territori és la fi que dóna sentit als treballs metodològics que aquí es presenten, tots ells incorporats a un Sistema d'Informació Geogràfica, com pot ser el SIG MiraMon, o dissenyats per a una propera incorporació. També el marc de treball és geogràfic des del punt de vista que els criteris cartogràfics i geodèsics, els aspectes d'anàlisi espacial, el models de dades i metadades, etc, han estat tractats i desenvolupats amb el màxim rigor geogràfic possible.

D'acord amb Hanging, geògraf del Department of Geography, University of Cambridge:

> *Geostatistical approaches to the analysis of spatial data have been underexploited in geographical research, particularly in human geography* (Hanging *et al.* 2010).

els temes abordats en aquesta tesi, que vol contribuir a una Geografia transversal i no compartimentada (Poon 2003; Unwin 1996) se situen en un camp encara no prou explorat i, per tant, aquest treball pot aportar una més gran col·laboració entre els dos àmbits: Geostadística i Geografia.

I, finalment, ens podem preguntar: i totes aquestes contribucions de la CIG per què serveixen? Una resposta la podem trobar en les reflexions del Committee on Strategic Directions for the Geographical Sciences in the Next Decade (2010):

> *The geographical sciences have the potential to improve understanding of the extent and causes of the changes unfolding on Earth's surface, to offer insight into the impacts of those changes, to promote the development of effective*

> *strategies in response to those changes, and to facilitate the documentation and representation of Earth's changing character.*

1.2 Ciència de la Informació Geogràfica

Un article recent (Reitsma 2013) reflexionava sobre si podem parlar de **Ciència** de la Informació Geogràfica, terme que va començar a fer fortuna amb Goodchild (1992) i que pocs anys després ja preocupava a Wright *et al.* (1997). En aquest treball creiem que sí, com hem reflectit en el seu títol, i considerem que les metodologies que són pròpies a la CIG són formalment comparables a altres ciències més clàssiques.

La Ciència de la Informació Geogràfica (CIG) es pot definir:

> *Un cuerpo de conocimiento que pretende el estudio, la investigación y el desarrollo de los conceptos teóricos, los algoritmos matemáticos, los programas informáticos, los instrumentos físicos, las bases de datos, las nuevas formas de uso y la búsqueda de nuevos campos de aplicación, en relación a las tecnologías de la información geográfica* (Bosque-Sendra 2005).

La figura 1, presa de Curran (2001), és un esquema/resum perfectament assumible per aquesta tesi en el context d'un punt d'encontre de les diferents disciplines de la CIG. També cal indicar que aquest esquema ressalta (no és un destacat nostre) dues de les disciplines més troncals en aquesta tesi: la Geostadística i la Teledetecció. A un segon nivell, caldria afegir la Geocomputació i, en un paper més funcional, però absolutament central, els Sistemes d'Informació Geogràfica (SIG). Altres disciplines com la Geodèsia, la Cartografia, etc, tenen en aquesta tesi un paper força més secundari.

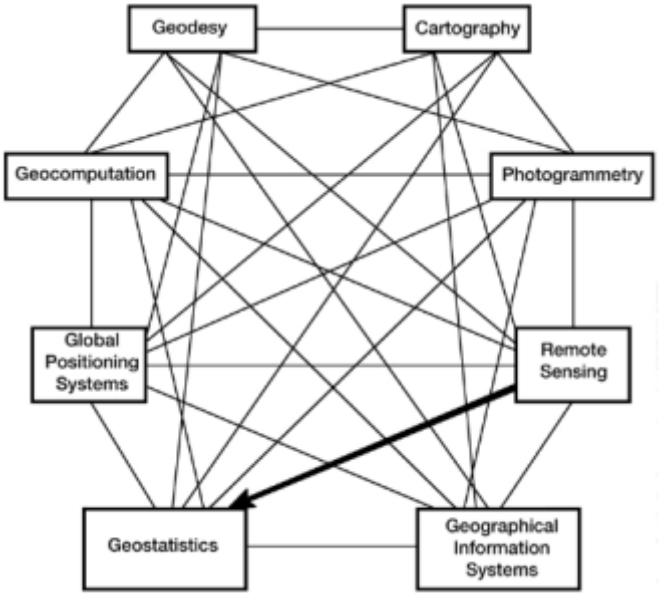


Figura 1: Connexions entre diverses disciplines de la Ciència de la Informació Geogràfica. Font Curran 2001.

En aquesta tesi, totes aquestes matèries tenen dos denominadors comuns, un de temàtic: l'**anàlisi espacial**, força lligat a la Geostadística en la majoria de capítols, i un de metodològic: l'**automatització**, molt enfocada a vessants computacionals.

Aquest denominador comú temàtic, l'anàlisi espacial, comprèn

> *els diferents conjunts de tècniques, geomètriques, topològiques, estadístiques o d'altre tipus, destinades a estudiar la localització i la distribució espacial dels fenòmens o entitats i la variació dels seus atributs temàtics en l'espai, així com les seves propietats i relacions espacials* (Nunes 2014).

I dins l'anàlisi espacial, aquesta tesi incorpora elevades dosis de Geografia Quantitativa:

> *Quantitative Geography consists en one or more of the following activities: analysis of numerical spatial data, the development of spatial theory, and the construction and testing of mathematical models of spatial processes* (Fotheringham *et al.* 2000)

Per tant podem resumir que, dins la CIG, aquesta tesi es focalitza principalment en la l'anàlisi espacial des d'un punt de vista quantitatiu, continuant la visió iniciada per Berry and Marble (1968).

1.2.1 El SIG

En aquest context, els Sistemes d'Informació Geogràfica són eines per a la Ciència de la Informació Geogràfica Goodchild (1992) o si ens referim a algunes definicions més explicatives:

> *A powerful set of tools for collecting, storing, retrieving at will transforming and displaying spatial data from the real world for a particular set of purposes* (Burrough and McDonnell 1998).

per la l'accepció que el descriu com un programari.

> *Sistema informatitzat de propòsit general per al maneig (captació, emmagatzematge, consulta, anàlisi, representació i difusió) d'informació geogràfica* (Pons and Arcalís 2012).

per la l'accepció més completa que engloba maquinari, programari, dades, protocols d'organització.

El SIG com a eina ha estat present de forma principal en tots els capítols d'aquesta tesi, i, a més a més, ha estat objectiu propi de recerca en el capítol 3. En el conjunt d'aquest treball, el rol funcional del SIG per a la selecció de les bases de dades i de les imatges adequades per a les diferents anàlisis, per a la georeferència precisa, per al tractament rigorós i complet de les metadades, etc, ha estat absolutament troncal. I en aquesta tesi, el SIG té un nom propi: ***MiraMon*** (figura 2a, figura 2b).
Com descriu el seu web de presentació:

> *El MiraMon es desenvolupa de forma cooperativa per part de diferents membres del Grup de Recerca Consolidat GRUMETS, pertanyents al Centre de Recerca Ecològica i Aplicacions Forestals (CREAF) i a la Universitat Autònoma de Barcelona (UAB); a data d'avui constitueix un SIG de propòsit general usat en*

> *àmbits científics, educatius i de gestió mediambiental que permet la visualització, consulta, edició i anàlisi tant de capes ràster, com vectorials, com geoserveis tipus WMS/WMTS. És utilitzat per unes 200 000 persones a 37 països.*

...i una d'aquestes 200 000 persones és l'autor d'aquesta tesi, que n'és usuari però també forma part del nucli desenvolupador.

Una de les principals aportacions estratègiques del MiraMon és l'aposta per facilitar dades (figura 2a) junt amb les eines (figura 2b) i alguna d'aquestes dades ha estat molt útil com a material auxiliar i de qualitat per determinats treballs: per exemple el Model Digital del Terreny de l'Institut Cartogràfic de Catalunya, completat amb l'ASTER Global Digital Elevation Model (Slater *et al.* 2011) a les zones transfrontereres.

Figura 2a: Capçalera de presentació de la versió 7 del SIG MiraMon, que mostra algunes de les col·leccions preferides de dades.

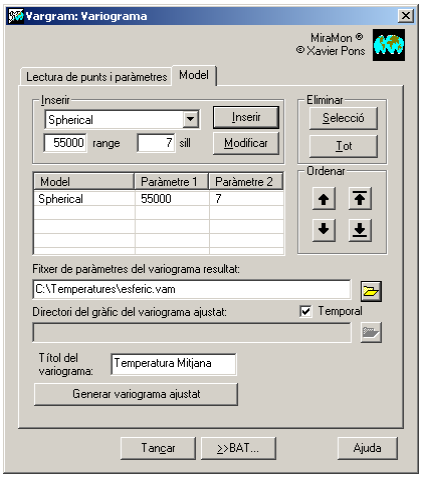


Figura 2b: Eina d'anàlisi del variograma del SIG MiraMon

Addicionalment cal considerar també com a aportació auxiliar d'aquesta tesi, la incorporació de metodologies avançades d'anàlisi espacial al SIG, en la línia que indica Griffith (2012):

> *Spawning from seeds of knowledge sowed by Cliff and Ord, parts of spatial statistics matured in the discipline of geography, and today promulgate through spatial analysis tools available in geographic information systems (GISs).*

1.2.2 La Teledetecció

La Teledetecció també ha estat objecte directe de recerca en aquesta tesi en el capítol 5 i en el capítol 6 (de fet, annex 1), cas d'ús principal en el capítol 4, i font de dades auxiliar per complementar els models analítics en el capítol 3. Ara bé, de forma coherent amb el caire transversal i integrador d'aquest treball, la Teledetecció sempre ha anat acompanyada del SIG i, com en altres treballs (Turner 2003; Ringrose *et al.* 1996) són les metodologies combinades de SIG i Teledetecció les que donen un valor afegit en aquesta recerca.

Podem definir la Teledetecció com:

> *Ciència i tecnologies que tenen per finalitat l'obtenció remota de dades a través de sensors, així com el seu processament i anàlisi aplicats a l'observació (des de l'espai, des de l'aire o des del terreny) de la Terra, de l'Univers, dels fons marins, etc.* (Pons and Arcalís 2012)

Aquesta tesi ha treballat amb diferents tipus d'imatges i sensors multispectrals i hiperspectrals, amb diferents nivells de processament, amb imatges d'alta i mitjana resolució, etc, però la principal font de dades protagonista ha estat la sèrie Landsat. Per molts motius, com per exemple pel fet que la resolució espacial dels seus sensors (MSS, TM i ETM+) és adequada per a l'anàlisi territorial dels patrons espacials dels àmbits d'estudi presents en aquesta tesi (secció 1.7), però també pel seu darrer gran èxit: la seva distribució gratuïta i sense traves burocràtiques (Woodcock *et al.* 2008) (serveixi aquest comentari per agrair a l'United States Geological Survey, USGS, la seva generosa distribució d'imatges).

Un resum de la tipologia d'imatges que s'ha usat en els diferents capítols d'aquest tesi és:

- Capitol3: Imatges Landsat-5 TM i Landsat-7 ETM+ per a generar índexs de vegetació (NDVI) i de terbolesa. Prèviament processades.
- Capítol4: Compressions amb pèrdua a diversos graus i anàlisis de patrons espacials sobre:
    - Imatges Landsat-5 TM: valors digitals sense georeferència precisa.
    - Imatges Terra MODIS: georeferenciades i en reflectàncies.
    - Imatges aeroportades CASI: valors digital i georeferenciades.
- Capítol 5: Imatges Terra MODIS, georeferenciades i en reflectàncies, amb ús addicional de màscares de qualitat.
- Capítol 6 (o Annex 1):
    - Imatges Terra MODIS georeferenciades i en reflectàncies.
    - Imatges Landsat 5-MSS, Landsat 5-TM (exemple a la figura 3) i Landsat 7-ETM+ per a recerca en metodologies de correcció radiomètrica.

Cal notar que només en el capítol 4 s'han usat les bandes espectrals corresponents als canals tèrmics dels sensors del Landsat que en presenten, mentre que en la resta d'estudis sempre s'ha treballat sobre les bandes de l'espectre solar, i que en cap capítol s'ha usat el canal pancromàtic del sensor ETM+.
Alhora, tant en la recerca com en el processament de tot aquest conjunt d'imatges,la Teledetecció ha proporcionat un banc de dades massiu per a comprovar el funcionament els processos computacionals, ja que majoritàriament s'ha treballat amb escenes completes com en el capítol 6 (per Landsat 5-TM, de 180 km aproximadament)·o amb subregions d'estudi extenses, que involucren un gran nombre de píxels (per exemple 562 500).

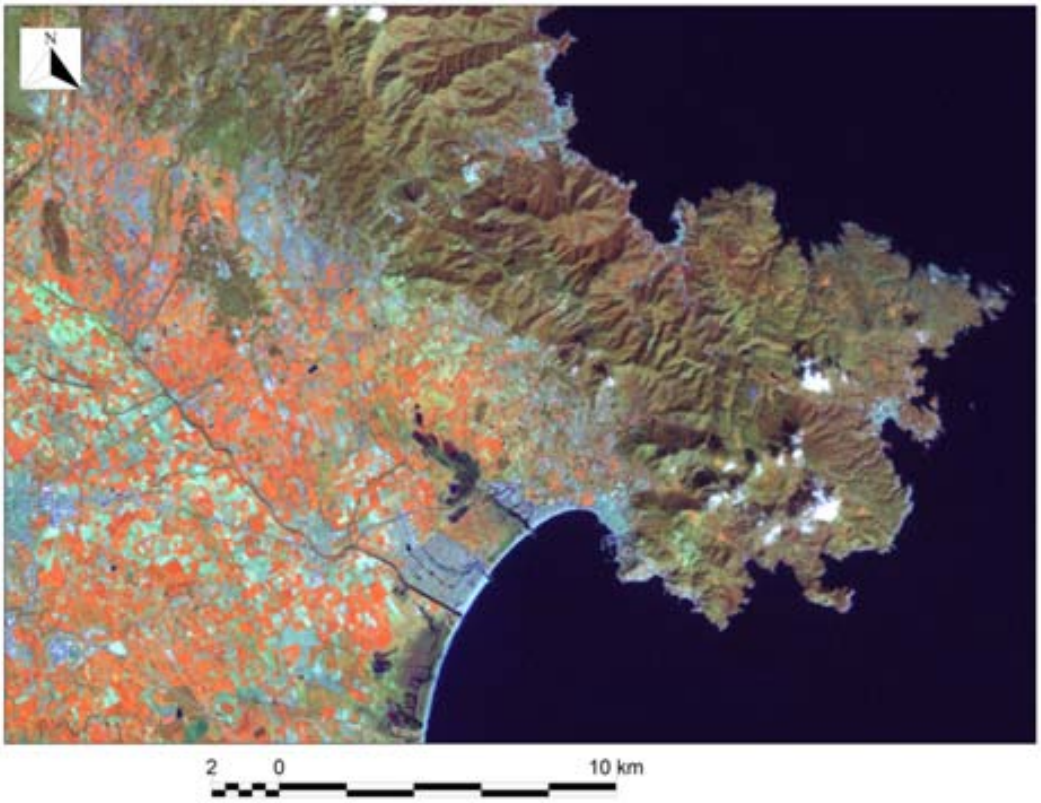


Figura 3: Composició RGB de les bades 4-5-3 d'una imatge Landsat-5 TM corresponent a l'escena del 10-3-2011, localitzada al cap de Creus. Font: Elaboració pròpia.

1.3 Geostadística

La Geostadística és una disciplina de l'Estadística especialitzada en l'anàlisi, la interpretació i la inferència de dades georeferenciades (Goovaerts 1997). Es basa en un conjunt de mètodes que permeten estudiar la distribució espacial d'una variable des d'una aproximació estadística, caracteritzar el seu patró espacial, estimar els corresponents valors a localitzacions problema, simular o modelitzar la seva variabilitat, etc.

La Geostadística s'ha aplicat habitualment sobre models de dades vectorials, majoritàriament corresponents a mostrejos en localitzacions puntuals, i aplicada a mètodes d'interpolació (Haining *et al.* 2010) i, en menys freqüència i més posteriorment, sobre models de dades ràster, essencialment sobre models digitals del terreny i imatges de Teledetecció (Curran and Atkinson 1998; Oliver *et al.* 2005, Balaguer-Beser *et al.* 2013). En aquesta tesi es presenten treballs geostadístics que il·lustren l'ús d'ambdós models de dades: Els capítols 2 i 3 corresponen a estudis sobre localitzacions puntuals que analitzen diferents tipus de variables: cotes altimètriques, dades climàtiques (temperatura mitjana mensual), freqüència d'espècies (flora i fauna), etc. En canvi, els capítols 4, 5 i 6 corresponen a anàlisis sobre models ràster, concretament sobre imatges de Teledetecció de diverses plataformes i sensors: Landsat MSS, TM i ETM+, Terra MODIS, i CASI (aeroportat). Tots ells tenen una característica comuna i innovadora en aquest camp, l'automatització, ja que habitualment (de fet, gairebé sempre) els

processos d'anàlisi geostadística es realitzen d'una forma força manual i interactiva (Jian *et al*.1996).

Dins de la Geostadística s'ha aprofundit de forma molt especial sobre el variograma i els paràmetres que defineixen la seva estructura: residu, amplitud, sostre i pendent (identificats com *nugget, range, sill* i *slope* en la bibliografia especialitzada, vegeu Kitanidis1997 i Cressie 1993 per la seva definició i interpretació) (figura 4).

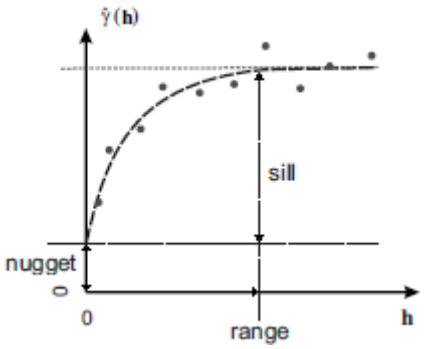


Figura 4a: Variograma teòric: paràmetres estructurals Font: Elaboració pròpia.

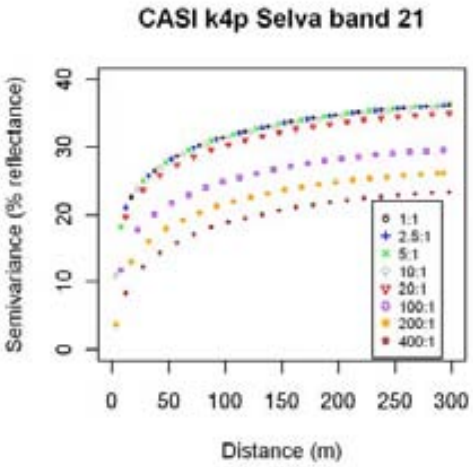


Figura 4b: Variograma empíric d'imatges de Teledetecció a diferents compressions. Font: Elaboració pròpia

En aquest treball hem automatitzat tant la construcció del variograma empíric com el seu ajust del teòric, bé sigui per permetre la seva incorporació a processaments d'interpolació distribuïts entre diferents unitats de processament, com per fer-ne un ús massiu. El variograma s'ha usat com a paràmetre de qualitat de les imatges comprimides i s'ha incorporat com a complement decisiu al test de detectors d'imatges d'elevada qualitat geomètrica i radiomètrica. Ens trobem doncs, davant d'un dels objectius de recerca més principals d'aquesta tesi.

1.4 Computació d'elevades prestacions

Gestionar grans volums de dades d'un forma eficient no és possible sense les eines, les metodologies i les infraestructures que aporta la Ciència Computacional. La computació d'elevades prestacions, identificada en la bibliografia especialitzada amb les sigles HPC (*High Performance Computing*) ha estat sovint més una eina (que per usar-la cal tenir certs coneixements de programació avançada) que un objectiu de recerca en aquest treball. Malgrat això, cal recalcar que els capítols 2 i 4 realitzen aportacions metodològiques prou rellevants (annex 3) i validacions en àmbits poc estudiats dins la Ciència computacional.

Les contribucions metodològiques que aporta aquesta tesi s'han testat en els laboratoris de recerca del Departament d'Arquitectura de Computadors i Sistemes Operatius

(DACSO) de la Universitat Autònoma de Barcelona, però no tenen cap adaptació forçada a les especificacions concretes de la seva infraestructura, sinó que s'han buscat solucions el més estàndard possibles perquè siguin vàlides en altres entorns. Per aquest motiu, s'ha optat per la implementació en llenguatge C del Message Passing Inteface (MPI), llibreria que permet la gestió coordinada d'un conjunt d'unitats de processament sense memòria compartida. El llenguatge C continua sent un dels llenguatges de programació més universals (King 2013) i MPI és un estàndard *de facto* (Walker *et al.* 1996), també molt usat en l'actualitat (Balaji *et al.* 2011) per a l'enviament d'informació mitjançant missatges que permet la distribució de càrrega computacional i la seva gestió organitzada.

L'accés a aquestes infraestructures d'elevades prestacions computacionals ha permès fer moltes proves, tot possibilitant canviar diferents paràmetres dels models, fer comprovacions amb diferents jocs de dades, introduir petites variacions, etc. També en fases avançades, però no definitives, dels diferents processaments i metodologies avaluades, quan semblava que els algorismes ja estaven consolidats, ha permès tornar a començar de nou, per tant obtenint resultats més contrastats, fiables i rigorosos. Obtenir aquests resultats en uns temps d'execució raonables malgrat treballar amb grans volums de dades, fa possible una anàlisi àgil i precisa dels diversos escenaris avaluats. Alhora, ha premés demostrar amb els paràmetres especialitzats d'eficiència computacional (p. ex. *speed-up*) que el disseny, la recerca i la solució de paral·lelització eren correctes.

1.5 Compressió d'imatges amb pèrdua

Les series temporals d'imatges multispectrals o hiperspectrals de Teledetecció tendeixen a acumular una gran quantitat d'informació geospacial, que en determinades situacions pot ser redundant o poc rellevant. Per exemple, podem trobar una important correlació entre diferents bandes espectrals, un temps de revisita excessiu o una resolució espacial massa detallada per a la dinàmica i patró espacial del paisatge de l'àmbit d'estudi. Llavors ens podem preguntar: És necessària tal quantitat d'informació? Fins a quin punt podem caracteritzar, analitzar, modelitzar, etc, el territori d'estudi amb menys volum d'informació?

Comprimir imatges amb pèrdua, o sigui, on ja no serà possible recuperar exactament la mateixa informació prèvia a la compressió, consisteix a rebutjar informació poc rellevant i guardar la més important (vegeu-ne un exemple a la figura 5). Però aquesta tria depèn fonamentalment dels usos que vulguem donar a les imatges (comprimides) i naturalment del tipus d'imatges i del territori i paisatge que està estudiant. En alguns casos i/o per alguns usos serà important conservar al màxim la dimensió espacial, mentre que en altres caldrà conservar especialment els detalls espectrals i/o temporals. Aquesta tesi inclou un capítol especialitzat en compressió d'imatges amb pèrdua, que alhora manté metodologies comunes als altres capítols de la tesi: anàlisi de patrons geostadístics i computació d'altres prestacions, aplicats altra vegada a imatges de Teledetecció.

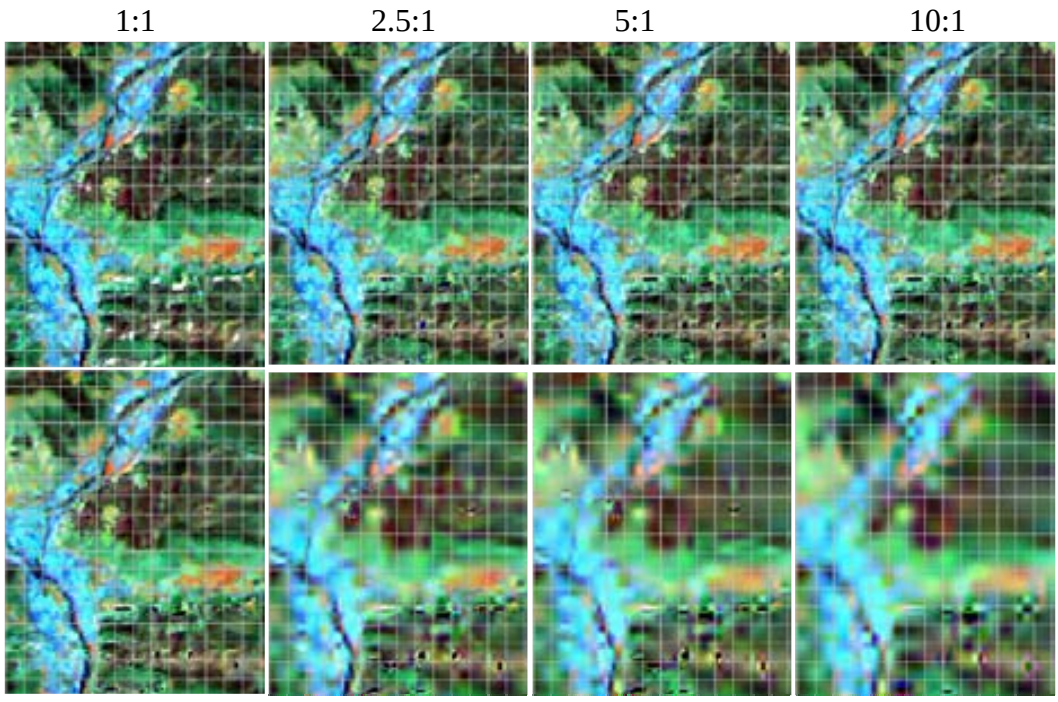


Figura 5: Imatge original i compressions amb pèrdua creixents, d'esquerra a dreta i de dalt a baix, d'una composició 4-5-3 com a RGB d'un retall a la comarca del Ripollès corresponent a l'escena 197-130 del 13-04-2006. La quadrícula, sobreposada per a facilitar la comparació, és de 200 m de costat. Font: elaboració pròpia.

En aquesta tesi essencialment s'han analitzat dues variants de la compressió d'imatges en l'especificació **JPEG2000**:

- BIFR (*band-independent fixed-rate*): compressió sobre les dues dimensions espacials sense tenir en compte cap altra informació.
- 3d-DWT(*three-dimensional discrete wavelet transform*): compressió que afegeix a les dues (X, Y) components espacials, una tercera, en alguns casos espectral (n bandes de la mateixa escena) i en alguns casos temporal (n dates per una mateixa banda espectral).

Les metodologies en compressió només han estat estudiades en un capítol, però la seva utilitat com a exemple de gestió optimitzada de grans volums de dades i les contribucions geostadístiques aconseguides aporten un cas d'ús força interessant al conjunt de la temàtica de la tesi.

1.6 *Big Data*

> *We have entered an era of Big Data. Through better analysis of the large volumes of data that are becoming available, there is the potential for making faster advances in many scientific disciplines and improving the profitability and success of many enterprises* (Agrawal *et al.* 2012).

Però, què entenem per *Big Data*? Una definició pràctica pot ser:

> *[…] too big, too fast,or too hard for existing tools to process.* (Madden 2012)

I una de més acadèmica:

> *Big Data concept means a datasets which continues to grow so much that it becomes difficult to manage it using existing database management concepts & tools. The difficulty can be related to data capture, storage, search, sharing, analytics and visualization etc.* (Singh *et al.* 2012)

Naturalment, el *Big Data* afecta molts camps i disciplines: Economia, Física, Medicina, Ecologia, etc, i les llargues sèries de dades geospacials també formen part d'aquesta allau d'informació digital que successivament va superant els nous prefixos que intenten quantificar la seva accelerada progressió (figura 6a).

```
1 Bit =  Binary Digit
8 Bits = 1 Byte
1024 Bytes =       1 Kilobyte
1024 Kilobytes =   1 Megabyte
1024 Megabytes =   1 Gigabyte
1024 Gigabytes =   1 Terabyte
1024 Terabytes =   1 Petabyte
1024 Petabytes =   1 Exabyte
1024 Exabytes =    1 Zettabyte
1024 Zettabytes =  1 Yottabyte
1024 Yottabytes =  1 Brontobyte
1024 Brontobytes = 1 Geopbyte
1024 Geopbytes =    1 Saganbyte
1024 Saganbytes =  1 Jotabyte
```

Figura 6a: unitats múltiples del byte.

Figura 6b: les 3 components del Big Data
Font: bigdatablog.emc.com/2013/08/08/mlb-a-big-fan-of-big-data

Els autors (Laney 2001; Singh *et al.* 2012) assenyalen que el concepte pot tenir 3 vessants: **volum** de dades (com en el cas "petapíxel" recollit a Pons and Arcalís 2012), **varietat** de dades i **velocitat** (figura 6b), i en aquesta tesi podem afirmar que en el seu conjunt s'ha abordat una notable varietat de dades i metadades (tot i que naturalment només geospacials) i un important volum; en aquest cas les dades massives provenen fonamentalment de la Teledetecció. El vessant de velocitat ha estat tractat de forma específica en el capítol 2 i en el capítol 4, lligat a la recerca en computació paral·lelitzada.

Aquesta tesi no és un projecte operacional que hagi tractat i analitzat un elevadíssim volum de dades, però com s'argumentarà en el capítol de resultats, les seves metodologies validades amb un significatiu volum de dades i, en alguns casos, en entorns computacionals complexos (HPC) sí que realitzen contribucions traslladables al *Big Data*, i concretament a un dels seus àmbits on s'obren més oportunitats per la recerca, el *Big Data Analysis* (Labrinidis and Jagadish 2012).

1.7 Àmbits geogràfics d'estudi

L'àmbit geogràfic d'estudi majoritari dels diferents treballs que es presenten en la tesi correspon a Catalunya, a diferents escales i extensions segons el cas. Només un treball està localitzat fora de Catalunya, en aquest cas a Andalucía (vegeu figura 7).

Figura 7: Mapa de situació dels àmbits d'estudi dels diferents capítols: Polígon vermell, Catalunya, àmbit de tots els capítols excepte del capítol 3, localitzat en una subzona (vegeu la figura 8) del polígon blau, Andalucía. Font: Elaboració pròpia.

El primer capítol comprèn dos estudis a àmbits i escales molt diferents: tot Catalunya a resolució espacial de 100 metres i un àmbit local a la ciutat de Barcelona de 380 m x 320 m a una resolució espacial de 0.5 m.

El segon capítol està localitzat al Parque Nacional de Doñana, a Andalucía (figura 8).

Figura 8: Superior. Situació (en verd) de l'àmbit d'estudi del capítol 3, Parque Nacional de Doñana, en la regió d'Andalucía (en rosa). Inferior: Detall del parc (gris-verd) i els seu entorn geogràfic. Font: Estación Biològica de Doñana.

El quart capítol compren diferents zones de Catalunya:, Ripollès, Penedès, Vallès i la Selva, a diferents resolucions espacials 3 m, 30 m i 250 m (en aquest darrer cas tota Catalunya) segons el tipus d'imatge de satèl·lit a analitzar.

El cinquè capítol i el sisè (o annex 1) tenen exactament el mateix àmbit que el *path-row* 197-031 del sistema WRS-2 de Landsat (Irish 2000) (figura 9).

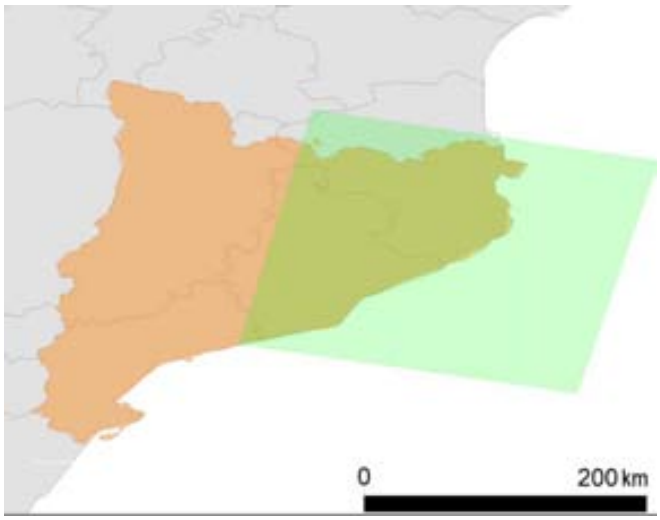


Figura 9: Intersecció del rectangle aproximat del *path-row* 197-31 del sistema WRS-2 (polígon verd) amb Catalunya (polígon taronja), àmbit d'estudi dels capítols 5 i 6. Font: Elaboració pròpia.

1.8 Justificació de la unitat temàtica de la tesi

Aquesta tesi es presenta com a compendi de publicacions i en la normativa de la Universitat Autònoma de Barcelona a la qual s'acull es demana que es justifiqui la seva unitat temàtica, aspecte que abordem a continuació.

En la introducció general i durant les seccions anteriors, que com s'ha vist estan naturalment molt entrellaçades, s'ha donat constants referències a la seva unitat temàtica. El resum específic per aquesta secció és que, malgrat no ser una tesi monotemàtica, és una tesi amb un grau molt elevat de coherència interna i externa. Interna perquè tots els capítols giren a l'entorn d'un element central temàtic: l'anàlisi espacial, i amb un element metodològic també central: la recerca en l'automatisme de processos. Dins de l'anàlisi espacial, la geostadística és l'actor principal amb components més propis dels Sistemes d'Informació Geogràfica en alguns capítols, i més pròpies de la Teledetecció en d'altres. Metodològicament, l'automatització és omnipresent a tots els capítols, sigui amb finalitats d'eficiència computacional, o com a via per al processament massiu de grans volums de dades.

Alhora, també voldríem ressaltar la seva coherència externa, o sigui respecte l'entorn de recerca més proper en què s'ha desenvolupat. El treball que es presenta és completament sinèrgic amb les línies d'investigació del grup de recerca del doctorand, Grumets (http://www.grumets.uab.cat/), reconegut per la Generalitat de Catalunya. Si s'analitzen els coautors de les diferents publicacions que configuren aquesta tesi, podem trobar dos fets que il·lustren aquesta coherència externa, alhora que posen en valor el seu caràcter de treball en equip, imprescindible en molts àmbits de la recerca actual:

- En totes les publicacions excepte en una, hi ha com a mínim un coautor del grup de recerca Grumets, a més a més del director de la tesi.
- Vuit diferents coautors de Grumets han participat en alguna de les publicacions i hi han participat sense repeticions (exceptuant naturalment el director).

Per tant, podem concloure que els capítols que segueixen a continuació formen una unitat temàtica. Alhora, les darreres seccions de resultats i conclusions globals, també reforçaran les interrelacions entres els diferents capítols.

Capítol 2

Parallel ordinary kriging interpolation incorporating automatic variogram fitting

Aquest capítol és una reproducció de Pesquer L., Cortés A., Pons X. (2011) “Parallel ordinary kriging interpolation incorporating automatic variogram fitting” Computers & Geosciences, 37, 464–473, doi:10.1016/j.cageo.2010.10.010.

Computers & Geosciences 37 (2011) 464–473

Contents lists available at ScienceDirect

Computers & Geosciences

journal homepage: www.elsevier.com/locate/cageo

Parallel ordinary kriging interpolation incorporating automatic variogram fitting

Lluís Pesquer [a,*], Ana Cortés [b], Xavier Pons [c,a]

[a] Center for Ecological Research and Forestry Applications, Edifici C, Univeristat Autònoma de Barcelona, E-08193 Bellaterra, Barcelona, Spain
[b] Computer Architecture and Operating Systems Department of Universitat Autònoma de Barcelona, Barcelona, Spain
[c] Geography Department of Universitat Autònoma de Barcelona, Barcelona, Spain

ARTICLE INFO

Article history:
Received 5 October 2009
Received in revised form
2 June 2010
Accepted 4 October 2010
Available online 13 November 2010

Keywords:
Kriging
Parallel programming
Variogram fitting
MPI

ABSTRACT

This work introduces a methodology for reducing the execution time of the kriging interpolation method without losing the quality of the model results, as occurs in simplified moving neighborhood solutions. The proposed solution distributes the computation applying parallel programming using MPI (Message Passing Interface) libraries in a HPC (High Performance Computing) environment. For the solution to be automatic and adaptable to different spatial patterns the variogram was automatically fitted; this preliminary modeling step is usually interactive in this interpolation method. The experimental results show the validity of the implemented solution, as it significantly reduces (in one of the examples the execution time decreases from 2 h 38 min to only 3 min) the final execution time of the entire process. The proposed solution is not exclusive to a particular architecture or operating system and can be applied in various environments and spatial resolutions of the generated raster model as well as at different magnitudes of the data to be interpolated.

1. Introduction

In some interpolation applications reducing the execution time can become the highest priority objective, for example in those applications that are part of support systems for quick decision making (predictions in risk situations, short-term meteorological forecasts, etc.). In this kind of application, we need to generate results as fast as possible in order to be useful for prediction purposes. Therefore, complex methodologies are often simplified or replaced by simpler (moving neighborhood) or faster (FFT based solution Fritz et al., 2009) methods, despite the interpolation model having lower prediction quality.

Computational Science offers methods for distributing computation across various processors with the aim of reducing the total execution time of processes with high computing needs; they therefore look for the best way to take advantage of the calculation capacity of high performance computing systems. In this field, parallel programming is one of the most efficient methodologies for achieving this objective and MPI (Message Passing Interface) is one of the most adaptable and powerful solutions. MPI is a standard implementation proposal for message passing programming paradigm in distributed systems (Message Passing Interface Forum, 1994; Snir et al., 1996) and has become the *de facto* standard message passing library, the successor of PVM (Parallel Virtual Machine; Gajraj et al., 1997; Geist et al., 1994).

Distributed computation in spatial analysis has distant antecedents (Openshaw 1987) as well as closer ones (Gajraj et al., 1997; Healey, 1996; Mineter and Dowers 1999); however, as Wang and Armstrong (2009) commented, it continues to be little used in this area. Spatial analysis methods are currently managing more and more information, and certain complex processes are becoming increasingly slower, as the hardware improves at a slower rate than the quantity of information for modeling increases.

Kriging, a geostatistical interpolation method, a discipline originated by Matheron (1962), is a paradigmatic case of a method whose computational cost increases greatly when it is applied to large amounts of data and predictive models of increasingly better planimetric resolution. Kerry and Hawick (1998), Lloyd (2006) and O'Sullivan and Unwin (2002) confirm its complexity and high computational needs. Kriging is a spatial interpolation option included in many Geographic Information Systems (GIS), such as *ArcGIS*, *GRASS*, *Idrisi* and *MiraMon*, and in some statistical software (*R*), surface analysis and representation software (*Surfer*), but it has few implementations for distributed environments.

This work aims to resolve this problem with a solution that maintains the predictive qualities of kriging but significantly reduces the execution time by means of parallel implementation. This proposed parallel kriging solution, unlike other solutions, is not specific to a particular environment or architecture such as that of Kerry and Hawick (1998) for *CM5*, or Gajraj et al. (1997) for

* Corresponding author. Tel.: +34 93 5811312; fax: +34 93 5814151.
E-mail address: l.pesquer@creaf.uab.cat (L. Pesquer).

0098-3004/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cageo.2010.10.010

Cray T3D. It does not use high level functions of specific libraries of a determined software, like for example Gebhardt (2003) and Rossini et al. (2003) for the API (Application Programming Interface) of the software *R*; these solutions are not directly portable or easily adapted to different configurations and new needs.

Moreover, the proposal presented here offers a methodological improvement: it does not assume a known spatial pattern, so that the process is carried out globally and also includes the spatial pattern modeling by means of automatic variogram fitting. To achieve this second objective a methodology inspired in Jian et al. (1996) has been applied with some differentiated elements.

The article first briefly reviews the most representative interpolation methodologies, essentially comparing their respective computational costs. The second section shows the algorithmic bases of the kriging method with the aim of better understanding the adopted parallel design and justifying the need to preserve the complete method, comparing it to certain simplified solutions that generate lower quality results. After this contextualization, the article focuses on detailing the design and implementation characteristics of the parallel programming procedure and showing the experimental results obtained with the output analyses tested and validated with heterogeneous data and environments.

2. Comparative computational analysis of the representative interpolation methods

There is extensive specialized bibliography that analyses and compares the different spatial interpolation methods; for example, Franke (1982) and Meyers (1994) represent two exhaustive studies in which comparisons between over 30 different methods can be found. It is also possible to find broadly summarized revisions with the most representative options in terms of purpose, such as Bonham-Carter (1994), Burrough and McDonnell (1998) and O'Sullivan and Unwin (2002), and in some cases these comparative analyses are approached from an application area (Kratzer et al., 2006). These more representative options usually have solutions implemented in the different software that include spatial analysis tools, and in particular in Geographic Information Systems (GIS; Cooper and Jarvis, 2004).

2.1. Inverse distance weighting (IDW)

This is a deterministic interpolation method in which the predicted value corresponds to a weighted average of the sample data points. The weight assigned to each sample depends, in an inversely proportional way, to the distance that separates it from the unsampled sites. The degree to which this dependency is modulated is the exponent and can be fitted in function of the chosen model. It can be formulated with Eq. (1).

Eq. (1): inverse of the weighted distance for the exponent β that modulates the distance r.

$$Z(x,y)=\frac{\sum_{i=1}^{n} z_i/r_i^{\beta}}{\sum_{i=1}^{n} 1/r_i^{\beta}} \quad r_i=\sqrt{(x-x_i)^2+(y-y_i)^2} \tag{1}$$

IDW is a method of low computational costs that depends on the number of point samples as well as prediction points (usually pixels), it is simple to implement and does not generate overflow.

Its characteristics are outlined in detail in Bartier and Keller (1996) and Shepard (1968).

2.2. Splines

There is a varied family of spline solutions (e.g. Eq. (2)), but generally they can be considered to be an interpolation method with models characterized by smooth forms without sudden changes in value that can predict maximums and minimums not present in the sample. This property, which is very interesting for certain applications, can result in unwanted or large overflow.

Their characteristics are outlined in depth in Mitasova and Mitas (1993).

Splines have a moderate calculation load that basically depends on the number of point samples and not on the number of unsampled sites (dimensions of the resulting model); therefore, a high number of point samples could imply large calculation costs.

Eq. (2): Splines proposed by Mitasova and Mitas 1993 composed by a pieces of radial functions with adjusted by φ tension, and T(x,y) flattening.

$$z(x,y)=T(x,y)+\sum_{i=1}^{n}\lambda_i R(r_i)$$
$$\rightarrow R(r_i)=-\left\{\ln\left[\frac{\varphi r_i}{2}\right]+E_1\left[\frac{\varphi r_i}{2}\right]+C_E\right\} \tag{2}$$

2.3. Kriging

Kriging is a geostatistical interpolation method of complex implementation, with certain difficulties involved in user learning and quite a large calculation load (Kerry and Hawick 1998; O'Sullivan and Unwin, 2002). As a method under study in this article, in the following sections we will analyze its foundations and more essential properties in more detail; however, a detailed review can be found in Cressie (1993).

Of the three representative methods, kriging has by far the highest computational cost, which is confirmed by references from the areas of spatial analysis (Lloyd, 2006), geostatistics (Pebesma and Wesseling, 1998) and HPC (High Performance Computing) (Gajraj et al., 1997; Kerry and Hawick, 1998). As these references are mainly qualitative, we found it necessary to carry out some tests that verify these differences. We carried out the interpolations, shown in Table 1, with a sample of 100 irregularly distributed points, generating an interpolated result in the form of a raster with 2673 columns $\times$ 2595 rows in the same computer and with 3 different software with the aim of analyzing the computational cost of the actual method and not a particular implementation. This set of tests corresponds to the parallel temperature interpolation, whose results are presented in Section 7.1.

The results confirm that kriging is much more computationally demanding than the other methods analyzed. Furthermore, the implementation is more complex, and as commented above, there are certain difficulties involved in user learning (Kitanidis, 1997).

3. Geostatistics

Kriging (or more precisely, ordinary kriging within the family of its geostatistical variants) estimates the value of the variable in a

Table 1
Execution times (seconds) comparing the 3 representative interpolation methods according to the implementation of 3 software with the data described in 7.1.

	IDW	Splines	Kriging
Idrisi	390	NA	6042
MiraMon	60	147	1347
Surfer	53	118	1467

localization problem from a weighted average of known values, as shown in Equation (3) (Oliver and Webster, 1990):

Eq. (3): predicted value in the problem localization as a weighted average of the n values that can be seen in the positions.

$$Z(\vec{x}_0) = \sum_{i=1}^{n} \lambda_i Z(\vec{x}_i) \tag{3}$$

The method for determining these weights is based on minimizing the variance of the variable, treated as a Regionalized Variable (RV), by modeling its variogram.

The RV theory provides the theoretical foundations for the basic operative tool of this interpolation method, the generation and modeling of the variogram, Eq. (4), which is analyzed in the next section. From this equation (Cressie 1993; Kitanidis 1997) we derive the relation (Eq. (5)) between the variogram model and the weights in Eq. (3), which will be applied to each sample to determine the prediction value.

Eq. (4): formula of the variogram.

$$\gamma(h) = \frac{1}{2} E\left[\{Z(x+h) - Z(x)\}^2\right] \quad \forall h \tag{4}$$

Eq. (5): equation for determining the weights that sample data points.

$$\sum_{i=1}^{n} \lambda_i \gamma(\vec{x}_i, \vec{x}_j) + \phi = \gamma(\vec{x}_j, \vec{x}_0) \quad \forall j \tag{5}$$

The computational analysis of the kriging interpolation algorithm showed that calculating the solution of the previous expression for each interpolated point (pixel) consumes a large part of the computational time, from approximately 96.0–99.8%, according to the example.

If Eq. (5) is translated to matrix notation, Eq. (6), the dimensions of the problem can be understood better.

Eq. (6): matrix notation of Eq. (5) and, by matrix algebra, deduction of the vector of weights.

$$\begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1n} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma_{n1} & \gamma_{n2} & \cdots & \gamma_{nn} & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \phi \end{bmatrix} = \begin{bmatrix} \gamma_{10} \\ \vdots \\ \gamma_{n0} \\ 1 \end{bmatrix}$$

$$A\lambda = b \rightarrow \lambda = A^{-1} b \tag{6}$$

4. Moving neighborhood

Before explaining the parallel programming solution proposed in this article, it is necessary to analyze the simplified proposal implemented widely in the software and used in many works: kriging with moving neighborhood. This solution basically consists in significantly reducing the dimensions of the matrices (Eq. (6)) that are used to obtain the weights that calculate the predicted value at each unsampled site, reducing the sample to a neighborhood close to this interpolated point (Hessami et al., 2001).

Reducing the dimensions of the problem achieves the objective of decreasing the execution time, but at the same time brings contradictions to the theoretical base, and practical problems that need to be considered.

The main problem from the methodological point of view stems from the fact that to carry out the interpolation the variogram is previously modeled and fitted for distances that are usually quite a bit larger than the moving neighborhood applied in this methodology. Therefore, if special care is not taken with this methodology (and none of the software analyzed have verification tools that help to maintain coherence between the range of the fitted variogram and that of the maximum neighborhood, nor did we find in the bibliography that authors that use this software explicitly confirm that the practical applications have been tested to maintain coherency), it can occur that only one zone of the variogram is being used when in fact it has been fitted for the overall area, the area for which the spatial structure has been identified.

Fig. 1. Discontinuities in the resulting model caused by the moving neighborhood method.

Secondly, it is necessary to add the effect of zonal interpolation and discontinuities that leads to the subsamples being different in close pixels (Fig. 1). This effect can be explained by the fact that an interpolated point collects a certain subsample in its particular neighborhood, and therefore the predicted value will be a weighted sum of the values of this subsample. It is therefore possible that a problem point corresponding to an adjacent cell for a moving neighborhood selects a slightly different subsample, leading to a significantly different predicted value.

Hence, the possible drawbacks of the moving neighborhood methodology mean that it is not always a valid solution for reducing the high computational cost of kriging. In other words and in agreement with Kitanidis (1997), when the calculation time cannot be assumed in a particular application it is more advisable to use one of the other lighter methods discussed in Section 2 than carry out a simplification of this methodology. Moreover, using all samples at each interpolation localization implies that is possible an efficient unique invert matrix step at Eq. (6) (Chilès and Delfiner 1999).

5. Variogram fit

Assuming therefore that the moving neighborhood technique is not a solution exempt from problems when the user wants to reduce the overall computational load, the solution proposed in this work is to distribute the load among different processors. If we start from the point samples and not from an already known or presupposed spatial pattern, constructing the variogram model is particularly important for distributing computation across different nodes.

This modeling is usually carried out interactively by the user for two reasons:

- An instrumental motive: the implementation in the great majority of software has a graphic interface that offers variogram analysis tools.

- An analytical motive: the interactive mechanisms oblige the user to analyze the spatial pattern of the sample and understand the proposed problem better.

Making the fit automatic deals with the instrumental motive and should be applied as a previous step to the parallel programming process;

In an analytical context, a standard user would use the steps described below:

5.1. Construction of the empirical variogram

1. Select the sample of data.
2. Choose the geometric variables of the variogram: lag width (Fig. 2) and number of intervals into which the range of distances to be analyzed is divided.
3. Analyze the possible anisotropy.
4. Visually analyze the variogram graphics in function of the chosen parameters in order to approximately identify the structural parameters of the variogram: nugget, range and sill (Fig. 2).
5. Analyze the statistical results of the variogram, as for example distances at each interval.

5.2. Generation of the variogram model

The aim is to find a function that best fits the points of the empirical variogram, as shown in Fig. 2.

Schematically, the user would carry out the following steps:

1. Choose the function that visually best fits the empirical variogram.
2. Determine the approximate parameters of the chosen function, which identify the aimed for structure: nugget effect, **slope** for the lineal model, **range** and **sill** for the rest.
3. Visually compare the variogram model to the empirical variogram.
4. Repeat point **3** until an appropriate solution is found. If the parameters chosen are close to an optimal solution, some software (e.g. *Idrisi* and *Surfer*) provide tools to finish fitting them, but this is not always possible.
5. Choose a different function and repeat steps 3 and 4.
6. Choose the model solution that user considers is most appropriate. The software provides tools that provide objective criteria for this choice.

In order to substitute an interactive procedure with an automatic method that is useful in a distributed systems, the method for determining the optimal overall solution proposed in this work takes into account the evaluations in Jian et al. (1996), who consider that the Least Squares (LS) fitting methods are clearly more efficient than the maximum likelihood (ML) methods. The LS implementation adopted has the following steps:

1. Separately fit the optimal solution for each of the implemented functions.
2. Collect the statistical criteria of each of the fits. The coefficient of determination R^2 (simple or adjusted) and χ^2 are the statistical criteria implemented.
3. Compare, for one statistical criterion, the different fits that minimize the error between the empirical model and the fitted model.

There are various possible least squares fits for point **1** for mainly non-lineal functions.

From the different variants proposed by Madsen et al. (2004), we chose the method by Levenberg–Marquardt (L–M), presented in the same reference (Madsen et al., 2004) and originally in Levenberg (1944) and in Marquardt (1963). For the coding we adapted the proposal by Press et al. (1988) to variogram model functions. Gaussian, exponential, spherical, linear, wave and quadratic are implemented model varigorams.

The L–M method is an iterative procedure, in which new values are calculated from initial values for incremental trials, which are progressively fitted to the appropriate values, from the partial derivatives of the function with respect to the parameters to be fitted, and which progressively come together, if possible, to a final solution.

In Table 2 we give an example of a specific fit (spherical model with nugget) using this method and the solution adopted in this work, specifying the new parameters at each iteration. A summary of this evolution (only iterations 1, 3, 5 and 7 are shown) can be graphically analyzed in Fig. 3.

It is important to point out that the automatic fit of the variogram is a quick procedure compared to the actual interpolation (it only consumes about 1–3% of the total execution time) and therefore it is not necessary to distribute its execution.

Table 2
Analysis of the evolution of the parameter fit with the interactive L–M method in the 7 first interactions.

Iteration	*Nugget*	*Range(m)*	*Sill*
1	0.07	40252.34	6.78
2	0.28	49223.43	6.34
3	0.25	52904.80	6.88
4	0.49	58621.89	6.83
5	0.47	58422.86	6.92
6	0.48	58650.98	6.89
7	0.41	58689.64	7.01

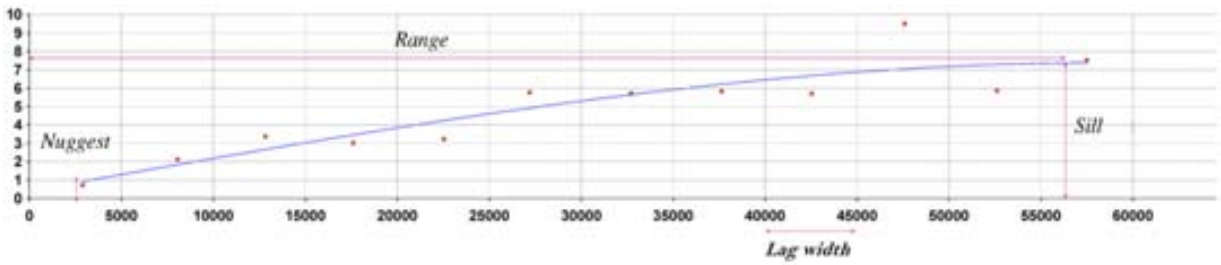


Fig. 2. Graphic of the empirical variogram (red dots) and the theoretical variogram (blue line) that allows the structural parameters to be identified of a variogram of 12 intervals 5000 m wide. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

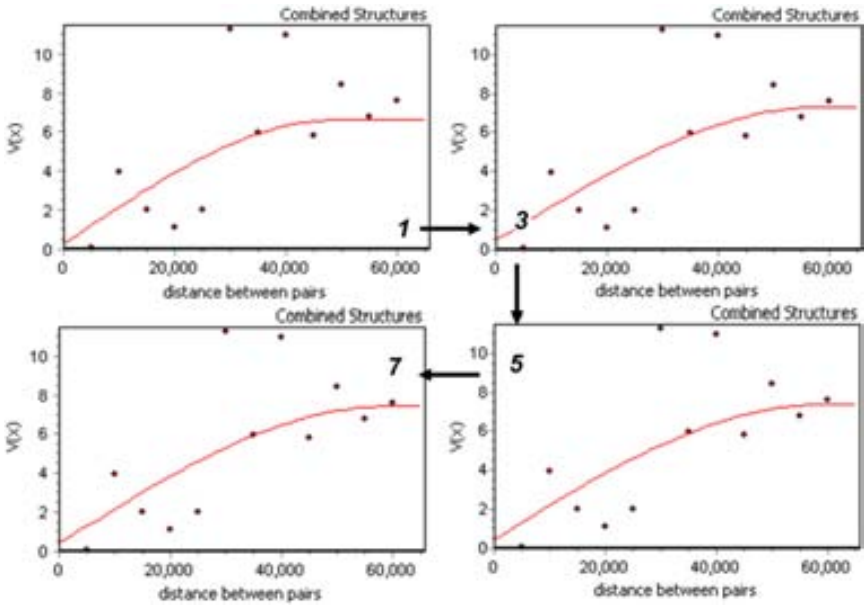


Fig. 3. Evolution of the fit of the initial values, iteration 1–7, according to the parameters in Table 2.

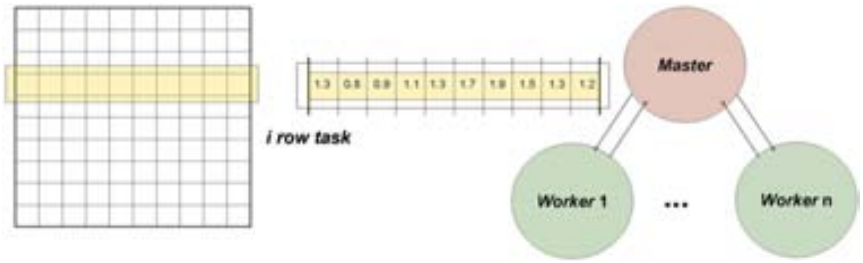


Fig. 4. Basic scheme of message passing in the master/worker design.

6. Parallel solution

From the analysis of the interpolation method detailed in Section 3 we have developed a parallel program by means of the MPI standard (Gropp et al., 1999; Pacheco, 1997). Below we outline its main characteristics.

The code adopted in the solution is written in ANSI-C language. This code has branched into preprocessor directives (#ifdef) to be able to commute between parallel executions that include the MPI library functions (version 1.2.7) and serial executions, which have been proved to be quite useful for removing errors from the implementation of the basic algorithm.

The computational analysis of the algorithm unmistakably indicates where it is necessary to distribute the workload. The solution adopted is the master/worker programming paradigm in which the master process is in charge of distributing and coordinating the tasks executed by the worker processes. These processes need to receive an adequate volume of tasks to execute, enough to compensate the additional time spent in communications but fragmented enough to provide flexible responses to the different possible characteristics of some of the interpolations and to be executed in different distributed environments.

In the implemented solution, the minimum data distribution unit (task) consists in a line of interpolated pixels (row of the raster to obtain).

The worker process hosts the core generation of the interpolation model, which consists in a row processing operator with the following basic parameters (Fig. 4):

- Receive a row identifier.
- Process the model for each cell of the row.
- Return a result line.

To process the model algorithm, the worker process (operator) first needs an information set that parameterizes the model according to the total dataset (invariant for the different rows):

- Total number of rows and columns and width of cells.
- Coordinates of the study area.
- Variogram model fitted automatically as detailed in Section 5.
- Data, values and position of the sample points.
- Global results, such as the inverted kriging matrix (only computed once).

As well as the resulting line, the operator returns certain addition values:

- The maximum and minimum of the line of values, which is useful for obtaining the maximum and minimum of the entire model; thus, correct information on the range of values can be given to the viewer software of the resulting digital model.
- The row identifier for building the ordered digital model.
- Possible error code to control the execution.

The task of the workers consists therefore in carrying out the line operator processes described above each time the master sends them a line.

The role of the master fundamentally consists of:

- Sending line processing orders to the different workers.
- Collecting the results.
- Saving the results on a disc.

These basic tasks require special care from the master in the following points:

- It is necessary to achieve the least possible time between receiving a row and sending a new one: it is important to take maximum advantage of the processing availability of the workers.
- It is necessary to correctly control the order of sending and receiving the rows (which can be different); the result will not make sense if it loses its original spatial structure.
- It is necessary to adequately control the progressive completion of the workers' tasks.

The master also carries out all the previous preparatory procedures; it is not necessary to distribute these procedures as the computation time is negligible, but it is necessary to remember to:

1. Read the data from the original format.
2. Construct the empirical variogram.
3. Fit the different variogram models implemented.
4. Select the best-fit variogram.

Points 2, 3 and 4 correspond to the process of automatic variogram fitting discussed in Section 5.

7. Experimental results

7.1. Interpolation of temperatures:

In this first validation of results, the variable to be interpolated corresponds to the average monthly temperature of the daily averages of the month of August 2005. These data were provided by the Catalan Meteorological Service (SMC) as part of a technological development agreement with the Center for Ecological Research and Forestry Applications (CREAF), which apply the methodological works of Ninyerola et al. (2000). These data come from the automatic network of SMC meteorological stations distributed throughout Catalonia, a region of approximately 32,000 km^2 situated to the northeast of the Iberian Peninsula, at the extreme southwest of Europe.

The spatial distribution of the stations can be seen in Fig. 5. Of the total sample, 100 fitting stations were chosen, which correspond to approximately 66% of the total, in order to reserve a subsample to validate the resulting model.

The sample chosen (the 100 stations; the rest were not involved in any way in the generation of the interpolated model) is characterized by the data given in Table 3, and the study area by the parameters shown in Table 4.

A descriptive characterization and comparison with the second dataset of the following section, summarized in Tables 3 and 4, shows that this sample is reduced, with an extensive area and that it generates an interpolated point matrix of considerable dimensions with a moderate spatial resolution (100 m). According to the number of stations and the study area, this is a too fine resolution for prediction maps but suitable for representing continuous maps. This implies a large number of estimation cells that offer an interesting very large computational problem.

Table 3
Statistical summary of the temperature data.

Number of data	100
Maximum (°C)	26.1
Minimum (°C)	9.5
Range (°C)	16.6
Mean (°C)	21.0
Std deviation (°C)	3.5

Table 4
Spatial characteristics of the temperature interpolation model.

Total area (km^2)	69,284.96
Pixel size (m)	100
Number of. columns	2672
Number of rows	2593
Number of pixels	6,928,496

Fig. 5. Map of the station locations (blue dots) of the automatic network of the SMC. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 5
Parameters of the automatically fitted variogram and of the quality of the fit in the temperature interpolation.

Nugget	0.44
Range (m)	58,463.64
Sill	6.94
χ^2	11.74
R^2	0.821
Adjusted R^2	0.754

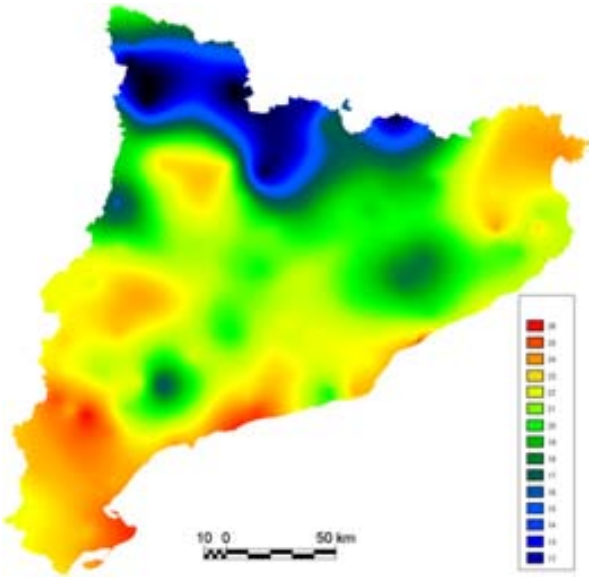


Fig. 6. Resulting raster of the interpolation of temperature dataset using variogram parameters of Table 2, 7th iteration.

Table 6
Table of computation times for the temperature interpolation in cluster 1. (*) In the master/worker design employed, the number of processors corresponds to the number of workers+1 (the master); therefore, 0 workers correspond to a single processor in a series procedure.

N Workers	Time (s)	Time (mm′ss″)
0*	1029.23	17′ 09″23
2	414.60	6′ 54″60
4	208.14	3′ 28″14
8	110.81	1′ 58″81
12	87.03	1′ 27″03

The empirical variogram, obtained from the parameters introduced as module arguments, which carry out the fit and interpolation, can be seen in Fig. 2 represented by red dots. This variogram automatically gives rise to the best-fit variogram model, represented by a blue line in Fig. 2, which in this case is a spherical model. The parameters and values of the statistical criteria that qualify the fit are detailed in Table 5.

The resulting raster of the interpolation, fitted after the study area, is shown in Fig. 6. This result is obtained identically in all the tests, which only differ in the execution time and never differ in model quality.

7.1.1. Cluster 1

The first tests in a distributed system were carried out on the cluster of the teaching laboratory of the Computer Architecture and Operating Systems Department of the Autonomous University of Barcelona. The cluster is composed by 12 nodes of Pentium IV 2.6 GHz processors each with 768 Mb of RAM, which communicate with an Ethernet network at 100 Mbps. Table 6 shows the averaged measures, and Fig. 7 shows the corresponding graphic representation and that of its output (speed-up, quotient between the serial time and the time of N processors).

7.1.2. Cluster 2

The second test environment was an integrated IBM cluster of the research laboratory of the same department, formed by 32 nodes each with two Dual Core Intel(R) Xeon (R) 3.0 GHz processors with 12 Gb of RAM, and communicated with Integrated dual Gigabit Ethernet. It has an overall capacity of 128 processors, and thus has more possibilities of analyzing in more depth the computational output of the proposed solution. Table 7 and

Table 7
Table of computation times for the temperature interpolation in cluster 2.

N Workers	Time (s)
0	80.0
2	43.4
4	24.6
8	15.3
12	12.1
16	11.3
20	10.1
24	10.4
24	10.3
28	8.9
32	13.9

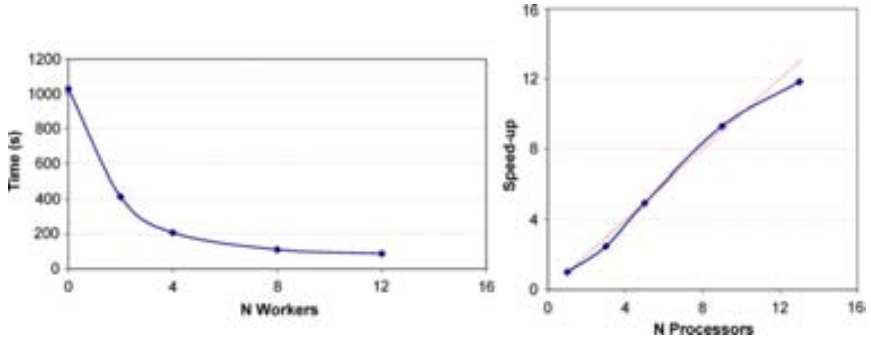


Fig. 7. Graph of the temperature interpolation output in cluster 1. On the left, execution time and on the right, optimum speed-up (red line) and empirical (points and blue line). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

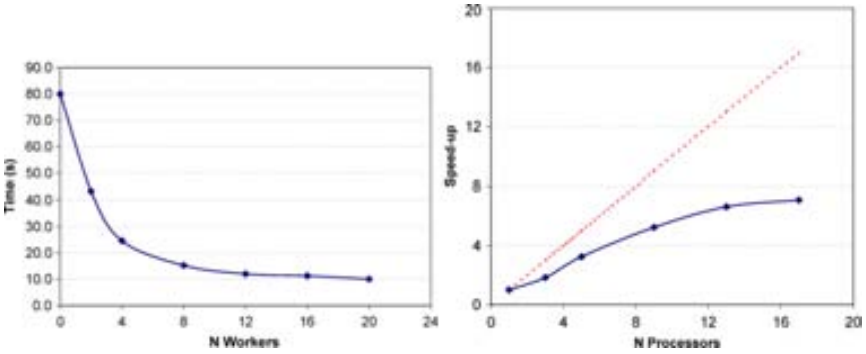


Fig. 8. Graph of the temperature interpolation output in cluster 2. On the left, execution time and on the right, optimum speed-up (red line) and empirical (points and blue line). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

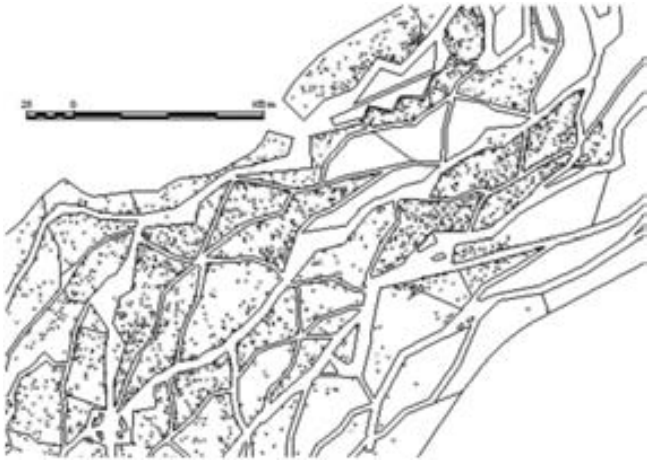


Fig. 9. Distribution of points with altitudes (black crosses) within the land blocks (outlined in black) of The Botanic Garden of Barcelona.

Table 8
Statistical summary of the altitude data.

Number of data	2855
Maximum (°C)	137.2
Minimum (°C)	101.2
Range (°C)	36.0
Mean (°C)	115.6
Std deviation (°C)	7.0

Fig. 8 show the computation time and output obtained in this cluster.

The two tests show interesting results: Cluster 2 achieves a shorter absolute time, but lower optimal efficiency. Cluster 1 has a relatively better output, basically because as it is less powerful the calculations are large enough to decrease the relative weight of the time invested in communications (sending data between nodes). This means that to improve the efficiency of cluster 2, it would be necessary to send larger packets of problem rows, which is a valid solution for nodes with enough memory resources such as cluster 2, but which would reduce the applicability of the proposed solution to more restricted environments.

7.2. *Interpolation of altitudes:*

The second validation test was carried out with the variable altitude with a subsample of three-dimensional position (planimetric

Table 9
Spatial characteristics of the interpolated model of altitudes.

Total area (m^2)	121,600
Pixel size (m)	0.5
Number of columns	760
Number of rows	640
Number of pixels	486,400

Table 10
Table of computation times for the altitude interpolation in cluster 1.

N Workers	Time (s)	Time (h m s)
0	41,172.00	11 h 26′ 12″00
2	22,559.37	6 h 15′ 59″37
4	9942.034	2 h 45′ 42″03
8	5373.00	1 h 29′ 33″00
12	4196.83	1 h 09′ 56″83
14	4699.19	1 h 18′ 19″19

Table 11
Table of computation times for the altitude interpolation in cluster 2.

N Workers	Time (s)	Time (h mm′ ss″)
0	9467.00	2 h 37′ 47″00
2	5211.90	1 h 26′ 51″87
4	3355.53	55′ 55″53
8	3011.03	50′ 11″03
12	1603.32	26′ 43″32
16	1226.76	20′ 26″76
20	1059.35	17′ 39″35
24	881.27	14′ 41″27
28	721.45	12′ 01″45
32	645.07	10′ 45″04
36	589.30	9′ 49″30
40	524.33	8′ 44″33
44	481.96	8′ 01″96
48	442.47	7′ 22″47
52	409.10	6′ 49″10
56	379.96	6′ 19″96
60	366.06	6′ 06″06
63	347.92	5′ 47″92
68	317.98	5′ 17″98
74	302.67	5′ 02″67
80	276.66	4′ 36″66
88	254.16	4′ 14″16
96	240.25	4′ 00″25
104	219.11	3′ 39″11
114	206.02	3′ 26″02
124	191.40	3′ 11″40
127	182.09	3′ 02″09

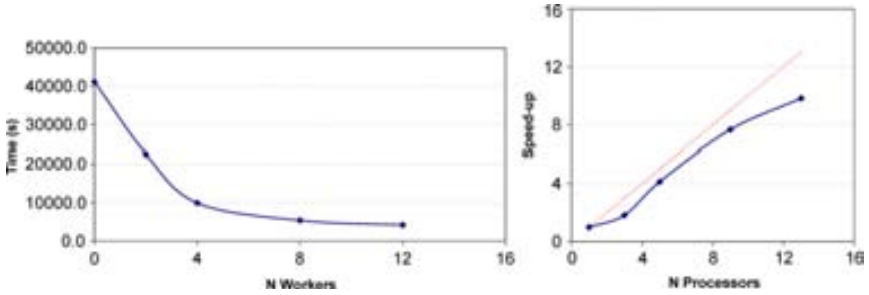


Fig. 10. Graph of the altitude interpolation output in cluster 1.

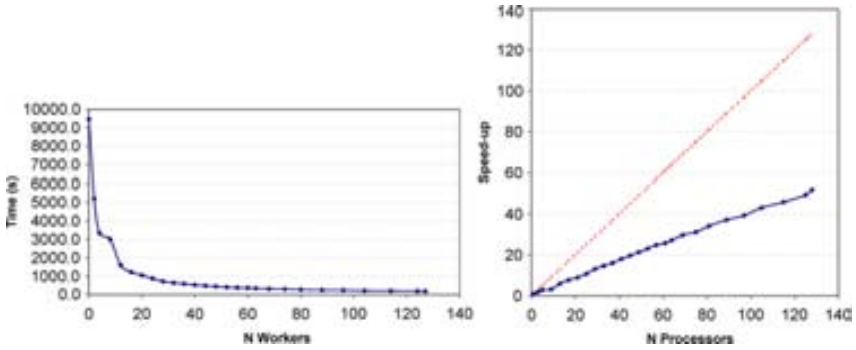


Fig. 11. Graph of the altitude interpolation output in cluster 2.

and altitudinal) measurements taken in The Botanic Garden of Barcelona (Spain).

The spatial distribution of the measures is shown in Fig. 9; their statistical values are shown in Table 8 and the study area in Table 9.

The corresponding results of the outputs of the two environments described above are shown in Tables 10 and 11 and Figs. 10 and 11.

The efficiency of cluster 1 is once again the most adequate, but in this case, as it is a far more demanding calculation, cluster 2 does not become totally saturated at its limit. Therefore, if the basic objective is to reduce the execution time, it can be seen that the proposed solution is useful and valid until the 128 processors of this cluster. It is also shown that the proposed solution is independent of a particular architecture or type of processor.

8. Conclusions

The solution proposed in this work for reducing the high computational cost of kriging does not follow the usual strategy of the simplification of this method at the price of losing predictive quality and then using a unique neighborhood with a complete dataset. This work focuses on maintaining the powerful theoretic framework of the original method and using a parallel programming implementation by means of MPI written in ANSI C language, standards which guarantee that the solution can be portable to different platforms.

To achieve the aim of a complete parallel interpolation process, it has been necessary to automatize the variogram fit and substitute the usual interactive procedure for carrying out this process. The solution of this fit is based on implementing the Levenberg–Marquardt method, an interactive procedure specifically for fitting non-lineal functions, on a set of variogram models and choosing with the aid of statistical criteria the optimal model from among those fitted.

The results obtained in two examples with very different characteristics (size of the sample and environment) show that parallel programming achieved the main objectives of significantly reducing the execution time without losing quality of the interpolated model and achieving a design that is flexible enough to be successfully adapted to different interpolations.

The experimental evaluation has been performed in two clusters with different characteristics what shows the portability of the proposed solution and allow us to contrast the validity of the solution in different computational environments.

The proposal discussed here is divided into two phases, variogram fit and parallel kriging interpolation algorithms, which are portable to other geostatistical interpolation methodologies that are similarly complex with high computational needs, as for example cokriging.

Acknowledgments

This research has been supported by the MICINN-Spain under contract TIN2007-64974. The authors also wish to thank to the anonymous reviewers for their constructive comments that helped to improve the manuscript.

References

Bartier, P.M., Keller, C.P., 1996. Multivariate interpolation to incorporate thematic surface data using inverse distance weighting (IDW). Computers & Geosciences 22 (7), 795–799.

Bonham-Carter, G.F., 1994. Geographic Information Systems for Geoscientists Modelling with GIS. Pergamon, p. 398.

Burrough, P.A., McDonnell, R.A., 1998. Principles of Geographical Information Systems. Oxford University Press, 333 pp.

Chilès, J.-P., Delfiner, P., 1999. Geostatistics: Modelling Spatial Uncertainty. Wiley, New York, 687 pp.

Cooper, W., Jarvis, C., 2004. A Java-base intelligent advisor for selecting a context-appropiate spatial interpolation algorithm. Computers & GeoSciences 30, 1003–1018.

Cressie, N.A.C., 1993. Statistics for Spatial Data, Wiley Series in Probability and Mathematical Statistics. Johm Wiley & Sons, New York, 900 pp.

Franke, R., 1982. Scattered data interpolation: tests of some methods. Mathematics of Computation 38, 181–199.

Fritz, J., Nowak, W., Neuweiler, I., 2009. Application of FFT-based algorithms for large-scale universal kriging problems. Mathematical Geosciences 41, 509–533.

Gajraj, A., Joubert, W., Jones, J., 1997. A Parallel Implementation of Kriging with a Trend Report LA-UR-97-2707, Los Alamos National Laboratory.

Gebhardt, A., 2003. PVM Kriging with R. In: Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vienna, Austria.
Geist, A., Beguelin, A., Dongarra, J., Manchek, R., Jiang, W., Sunderam, V., 1994. PVM: A Users' Guide and Tutorial for Networked Parallel Computing. MIT Press, p. 176.
Gropp, W., Lusk, E., Skjellum, A., 1999. Using MPI: Portable Parallel Programming with the Message-Passing Interface. MIT Press, 368 pp.
Healey, R.G. (Ed.), International Journal of Geographical Information Systems, 10; 1996, pp. 667–668.
Hessami, M., Anctil, F., Viau, A.A., 2001. Delaunay implementation to improve kriging computing efficiency. Computers & Geosciences 27, 237–240.
Jian, X., Olea, R.A., Yu, Y., 1996. Semivariogram modeling by weighted least squares. Computers & Geosciences 22–4, 387–397.
Kerry, K.E., Hawick, K.A., 1998. Kriging Interpolation on High-Performance Computers, Technical Report DHPC-035 Department of Computer Science, University of Adelaide, Australia.
Kitanidis, P.K., 1997. Introduction to Geostatistics: Applications to Hydrogeology. Cambridge University Press 249 pp.
Kratzer, J.F., Hayes, D.B., Thompson, B.E., 2006. Methods for interpolating stream width, depth, and current velocity. Ecological Modelling 196, 256–264.
Levenberg, K., 1944. A method for the solution of certain non-linear problems in least squares. Quarterly Applied Mathematics 2 (2), 164–168.
Lloyd, C.D., 2006. Local Models for Spatial Analysis. CRC Press, Belfast, 244 pp.
Madsen, K., Nielsen, H.B., Tingleff, O., 2004. Methods for non-linear least squares problems, Informatics and Mathematical Modelling. Technical University of Denmark, 58 pp.
Marquardt, D.E., 1963. An algorithm for least squares estimation of nonlinear parameters. Journal of the Society for Industrial and Applied Mathematics 11 (2), 431–441.
Meyers, D.E. (Ed.), 1994. Spatial Interpolation: An Overview. Geoderma (62), pp. 17–231 (special issue).
Matheron, G., 1962. Traité de Géostatistique Appliquée. Editions Technip, Paris, 334 pp.
Message Passing Interface Forum, 1994. MPI: a message-passing interface standard. International Journal of Supercomputer Applications 8, 165–414.
Mineter, M.J., Dowers, S., 1999. Parallel processing for geographical applications: a layered approach. Journal of Geographical Systems 1 (61), 74.
Mitasova, H., Mitas, L., 1993. Interpolation by regularized spline with tension. Mathematical Geology 25 (6), 641–655.
Ninyerola, M., Pons, X., Roure, J.M., 2000. A methodological approach of climatological modelling of air temperature and precipitation through GIS techniques. International Journal of Climatology 20, 1823–1841.
Oliver, M.A., Webster, R., 1990. Kriging: a method of interpolation for geographical information systems. International Journal of Geographical Information Science 4 (3), 313–332.
Openshaw, S., 1987. Some applications of supercomputers in urban and regional analysis and modelling. Environment & Planning A 19, 853–860.
O'Sullivan, D., Unwin, D., 2002. Geographic Information Analysis. John Wiley & Sons, Hoboken NewJersey, 436 pp.
Pacheco, P.S., 1997. Parallel Programming with MPI. Morgan Kaufman Publishers, 418 pp.
Pebesma, E.J., Wesseling, C.G., 1998. Gstat, a program for geostatistical modelling, prediction and simulation. Computers & Geosciences 24 (1), 17–31.
Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., 1988. Numerical Recipes in C. Cambridge University, New York, 963 pp.
Rossini, A., Tierney L., Li N., 2003. Simple Parallel Statistical Computing in R. UW Biostatistics Working Paper Series University of Washington, 93.
Shepard, D. 1968. A two-dimensional interpolation function for irregularly-spaced data. In: Proceedings of the 1968 ACM National Conference. pp. 517–524.
Snir, M., Otto, S., Huss-Lederman, S., Walker, D., Dongarra, J., 1996. The MPI Complete Reference Massachusetts Institute of Technology Press, 352 pp.
Wang, S., Armstrong, M.P., 2009. A theoretical approach to the use of cyberinfrastructure in geographical analysis. International Journal of Geographical Information Science 23 (2), 169–193.

Capítol 3

Automatic modelling and continuous map generation from georeferenced species census data in an interoperable GIS environment

Aquest capítol és una reproducció de Pesquer L., Prat E., Díaz-Delgado R., Masó J., Bustamante J., Pons X. (2012) "Automatic modelling and continuous map generation from georeferenced species census data in an interoperable GIS environment" com a capítol de Seppelt R., Voinov A.A., Lange S., Bankamp D. (Eds.) (2012):*International Environmental Modelling and Software Society (iEMSs)* 2012 International Congress on Environmental Modelling and Software Managing Resources of a Limited Planet: Pathways and Visions under Uncertainty, Sixth Biennial Meeting, Leipzig, Germany. http://www.iemss.org/society/index.php/iemss-2012-proceedings. ISBN: 978-88-9035-742-8.

International Environmental Modelling and Software Society (iEMSs)
2012 International Congress on Environmental Modelling and Software
Managing Resources of a Limited Planet, Sixth Biennial Meeting, Leipzig, Germany
R. Seppelt, A.A. Voinov, S. Lange, D. Bankamp (Eds.)
http://www.iemss.org/society/index.php/iemss-2012-proceedings

Automatic modelling and continuous map generation from georeferenced species census data in an interoperable GIS environment

Lluís Pesquer [a], Ester Prat [a], Ricardo Díaz-Delgado [b], Joan Masó [a], Javier Bustamante [b] and Xavier Pons [c]
[a] CREAF, Cerdanyola del Vallès 08193, Spain. l.pesquer@creaf.uab.cat
[b] Remote Sensing and GIS Lab (LAST-EBD). Estación Biológica de Doñana, CSIC. rdiaz@ebd.csic.es
[c] Departament of Geography, Universitat Autònoma de Barcelona. xavier.pons@uab.cat

Abstract: The Natural Processes Monitoring Team from the Doñana Biological Station (EBD), systematically acquires data on more than 100 indicators of ecological processes and the status of many fauna and flora species in Doñana National Park, one of the most important protected wetlands in Europe, covering 54000 Ha. This information is available on a website as tabular data and trend charts. A detailed analysis is necessary in order to interpret this information and provide decision-making criteria for the management of the natural area.
The purpose of this paper is to improve public access to the information collected in the monitoring program and at the same time increase its quality. The proposed methodology integrates spatial interpolation methods, multivariate linear and logistic regression models (including the use of remote sensing images as predictors) and hybrid tools of these methodologies into a Geographic Information System (GIS) based model to generate predictive maps of ecological parameters. The information on the distribution, abundance, population structure and densities of different terrestrial and aquatic species, biophysical parameters and also their corresponding validation methodologies were used in the automatic generation of continuous maps of the distribution and abundance of the species in the study region.

***Keywords**:* GIS; automatic modelling, interoperable web map.

1 INTRODUCTION

Public administrations, institutions and research centres are currently devoting considerable effort to projects that collect large datasets of environmental geoinformation. For example, the Natural Processes Monitoring Team of the Doñana Biological Station (EBD) systematically acquires information on more than 100 indicators of ecological processes and on the status of many fauna and flora species (Díaz-Delgado [2010]). This information is available on a website in the form of tabular data and trend charts. Therefore, specific analyses are needed to interpret this information (Bonham-Carter [1994]) so that it can be used to provide decision-making criteria for research and management purposes (Scotts and Drielsma [2003]). A continuous map representation of the spatial distribution of these datasets would increase the number of potential users and facilitate analytical applications, and thus these public resources would be used more effectively.

The present work describes a methodology for automatically generating maps from species census data, which are then published on a web map server. Chain process automation, which includes acquiring and filtering data, map generation and web map publication, requires a well-analysed design and suitable tools that include self-decision procedures. Metadata in this context have an important function, and only with accurate knowledge of the quality indicators of the different steps of the chain process allows making automatic decisions.
Some previous works (Kiehle *et al.* [2007], Walter *et al.* [2011]) have emphasized the key role played by metadata in web map services, and here, in the present work, metadata plays an additional consequential role in the steps before web publication: the metadata are used to improve the models as well as map generation. The models and methods selected here have already been applied in similar environments (Valley *et al.* [2005], Hancock and Hutchinson [2006]). The aim of this work, however, is not to determine an optimal prediction method but rather to design and implement an entire automatic solution integrated into a GIS environment following standard and interoperable protocols.

2 METHODOLOGY

The proposed methodology is a chain of automatic processes that goes from downloading data on different servers to adding the corresponding map to a web portal for its publication. Figure 1 summarizes a flowchart of the entire chain process.

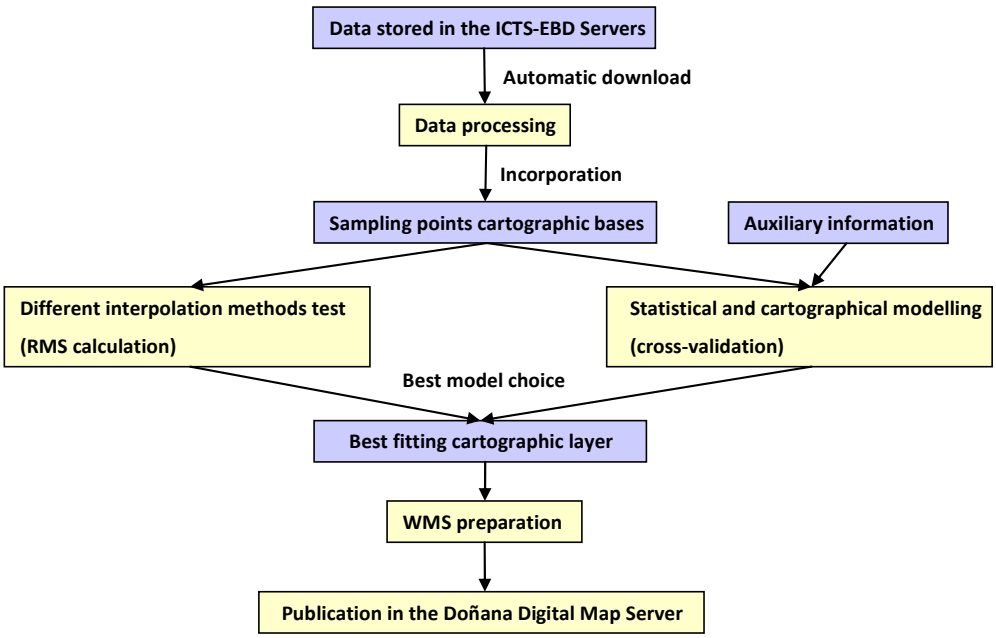


Figure 1. Flowchart of the proposed methodology.

All of these processes are integrated in the MiraMon GIS software (Pons [2000]), which allows chaining and controlling the flow execution by metaprogramming commands in a BATCH Windows environment (Microsoft [2010]). The main characteristic of this software is that the metadata are managed accurately, which allows, at every partial result, to program automatic decisions based on quality information from the previous step in the process chain.
It is important to note that the species and models were chosen for map generation solely in order to demonstrate the feasibility of automating the entire procedure, and the objective was not to obtain the most suitable inference method for each species. This work focuses on automatic decisions that depend on a specific dataset and are based on the comparisons of the results of a large number of tests.

The study region is the Doñana National Park, one of Europe's most important wetland reserves (Díaz-Delgado [2010]) and a major site for migrating birds. It covers an area of over 54000 Ha and is located in the south of the Iberian Peninsula, southwestern Europe (Figure 2).

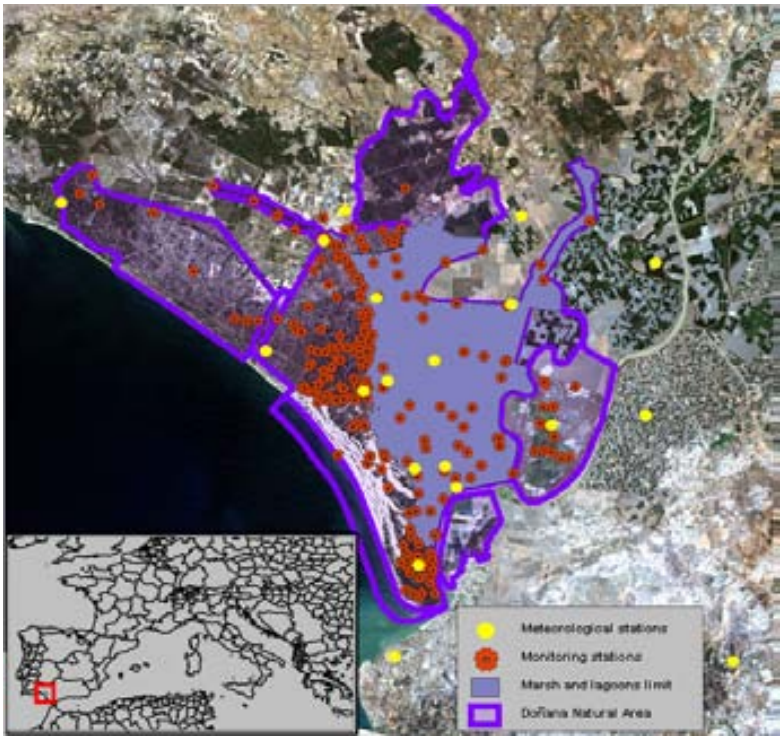


Figure 2. Location of Doñana National Park and the sampling stations.

2.1 Materials and data processing

Four different types of data are downloaded:

- Meteorological variables, such as mean air temperature, maximum air temperature, minimum air temperature and precipitation;
- Abundance of aquatic fauna, for example, the number of Lousiana crayfish (*Procambarus clarkii*) individuals;
- Presence/absence of aquatic vegetation; for example, the salt-marsh bulrush (*Boloboschoenus maritimus)*;
- Hydrological variables, such as water level, water temperature, minimum water temperature and maximum water temperature.

Meteorological variables are obtained from two different sites:

- Automatic weather stations from the Singular Scientific and Technological Infrastructure (ICTS) of the Doñana Biological Reserve: http://icts.ebd.csic.es/GeneradorDatosXMLGeneralServlet.
- Agrometeorological stations of the Research and Training Institute for Agriculture and Fisheries (IFAPA): http://www.juntadeandalucia.es/agriculturaypesca/ifapa/ria/servlet/FrontController .

Hydrological variables are also downloaded from the ICTS website at http://icts.ebd.csic.es/GeneradorDatosXMLGeneralServlet. The ICTS server provides data in XML format (Gutiérrez et al. [2003]), as shown in Figure 3, while the IFAPA server provides data in plain text format. These datasets require several post-processing procedures to prepare them for modelling, such as detecting and erasing wrong data, fusing coherently from different origins, grouping and average calculating.

```
▼<equipos>
  ▼<equipo id="2" nombre="MANECORRO RM1" fechaUltimaObservacion="3/02/12
   19:40" fechaInicioTrabajo="19/03/08 0:00" umtx="189502.46"
   umty="4114242.53" estado="0">
     <tipoEquipo id="31" nombre="Estación Meteorológica" frecuencia="10"
     numVars="21" modelo="VAISALA WTX 510"/>
    ▼<observaciones numero="3024">
       <observacion id="19571867" fecha="4/07/09 1:00"
       nombrevariable="Dirección del viento máxima" valor="335.0"
       calidad="Correcto" unidades="grados"/>
       <observacion id="19571866" fecha="4/07/09 1:00"
       nombrevariable="Dirección del viento media" valor="284.95"
       calidad="Correcto" unidades="grados"/>
       <observacion id="19571865" fecha="4/07/09 1:00"
       nombrevariable="Dirección del viento mínima" valor="253.0"
       calidad="Correcto" unidades="grados"/>
```


Figure 3. Example of meteorological variables in an XML file from the ICTS server http://icts.ebd.csic.es/GeneradorDatosXMLGeneralServlet?idEstacion=3&fechaInicio=040720090000&fechaFin=040720092359

For the data on aquatic fauna abundance and the presence/absence of aquatic vegetation, the download site is http://icts.ebd.csic.es/GeneradorDatosSeguimientoXMLServlet and similar processes of filtering and aggregating data are involved.

2.2 Map and model generation

Previous studies have tested different methodologies for generating a continuous representation of a quantitative variable from the values observed in specific locations (Lloyd [2006]). The prediction and modelling methods implemented were selected from the most usual methods (Burrough and McDonnell [1998]) and those that are most suitable for automation, for example kriging interpolation was discarded due to the difficulty involved in obtaining an automatic variogram (Pesquer *et al.* [2011]). Continuous map generation methods include different kinds of strategies: univariate and multivariate, and based on spatial patterns or statistical regressions, among others.
The specific methodologies tested were:

- Two spatial interpolation methods, inverse distance weighting (IDW) (Bartier and Keller [1996]) and splines (Mitasova and Mitas [1993]) used for meteorological and hydrological variables and for aquatic fauna abundance.
- Logistic regression (Kleinbaun [1994]) to determine the probability of the presence of aquatic vegetation.
- Multivariate regression + residual spatial interpolation (Ninyerola *et al.* [2000]), for meteorological variables and aquatic fauna abundance.

The final map is chosen by comparing the accuracy of the results generated by using different parameter sets for the same method or comparing the results of different methods. This estimate of the accuracy is obtained with an independent test validation subset or by using a leave-one-out cross-validation (Isaaks and Srivastava [1989]) for small samples. In addition, these validations and comparisons are integrated automatically into the entire flowchart.
The maps generated have a spatial resolution of 150 m, which is based on the distribution of the sampling locations, (Hengl [2006]) and they are georeferenced in the UTM-29N reference system.

2.3 Web publication

Publishing continuous maps for the different variables studied in a web environment is an essential step for the widest possible dissemination of results. Furthermore, using OGC standard protocols (Open Geospatial Consortium [2008]), such as WMS map servers, makes it possible to integrate data generated in the

emerging Spatial Data Infrastructures and provides interoperability with other available information.
The automated integration of maps into the existing Doñana Biological Station server (http://mercurio.ebd.csic.es/seguimiento) is the final step in the project. The CreaMMS tool (Maso and Pons [2005]) from the MiraMon software makes it possible to prepare layers that will be served later, and which will therefore be accessible to any OGC client.

3 RESULTS

3.1 Maps and models

A collection of continuous maps for each annual hydrological cycle (from September to August) and within the period 2007-2011 have been obtained. The maps include meteorological variables, hydrological variables, abundance of aquatic fauna and the probability of the presence of aquatic vegetation.
Three representative examples are shown below:

a) Spatial interpolation of precipitation (2010-2011).
Table 1 compares different quality results (by RMS) at different exponent parameter values for the IDW method. A range of tension parameter values have been used in the splines method.

IDW		Splines	
Exp.	**RMS**	**Ten.**	**RMS**
1	202.71	25	332.25
1.25	203.99	50	201.08
1.50	206.04	75	190.53
1.75	208.59	100	189.12
2.00	211.39	***125***	***189.09***
2,25	214.23	150	189.19
2.50	217.02	175	189.28
2.75	219.71	200	189.37

Table 1: RMS for each exploration parameter: exponent for IDW and tension for splines.

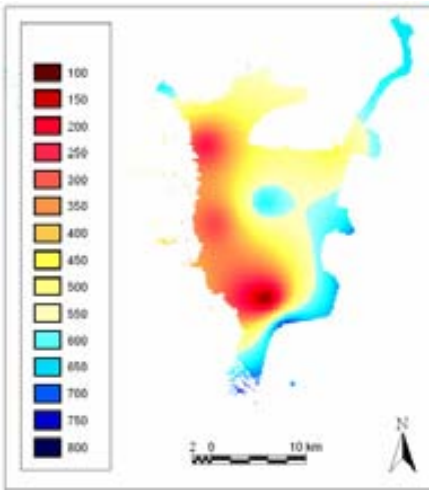


Figure 4: Spline interpolation for precipitation in the period 2010-2011.

In this example, most spline interpolations are better than the best IDW. A spline tension of 125 was selected for map generation (see Figure 4) as it has a small RMS.

b) Multivariate linear regression and residual spatial interpolation for Lousiana crayfish abundance (2007-2008).
Three variables were initially entered into the regression model in this example: hydroperiod (number of rainy days during a complete cycle) was automatically excluded, and the probability of the presence of aquatic vegetation and the average maximum temperature were selected, as shown in Table 2. This regression generated residual values for sampling locations were spatially interpolated until a minimum RMS was obtained. Figure 5 shows the final result, the regression model + spatial interpolation of the regression residuals.

c) The probability of the presence of salt-marsh bulrush (2007-2008) using logistic regression.
Table 3 provides a case study that models the probability of the presence of an aquatic species. Three remote sensing products were introduced as possible

auxiliary variables: the maximum and average NDVI for the entire period (Rouse *et al*. [1973]) and the marsh water turbidity (Bustamante *et al.* [2009]).

Independent Variables	Signi-ficant	Coefi-cient
Hydroperiod	No	
Aquatic vegetation probability	Yes	227.288
Average maximum temperature	Yes	49.219
Intercept		1235.588

Table 2: Independent variables introduced into the linear regression model and corresponding results.

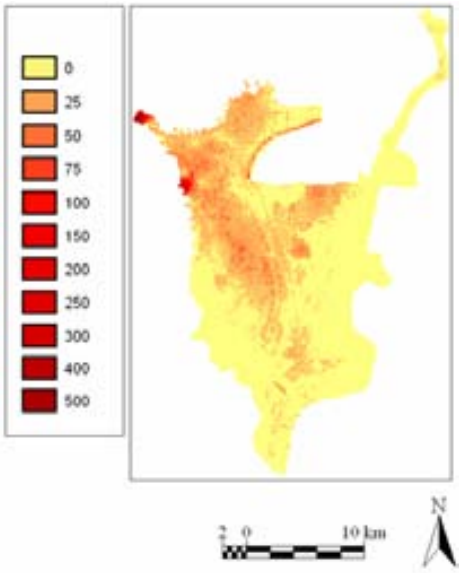


Figure 5: Map of the regression model + residual interpolation for *Procambarus clarkii* abundance

Two meteorological variables were used, hydroperiod and mean maximum temperature for the complete period, and finally the average distance to the flooded area was also introduced (Díaz-Delgado *et al*. [2006]). In this example hydroperiod and maximum NDVI were not significant and were not used in the final regression model. Figure 6 shows the resulting map generated from the Table 3 regression.

Independent Variables	Signifi-cant	Coefi-cient
Hydroperiod	No	-
Maximum NDVI	No	-
Turbidity	Yes	-0.001
Average distance to the flooded area	Yes	0.001
Average NDVI	Yes	-4.842
Average of maximum temperature	Yes	1.327
Intercept		32.715

Table 3: Independent variables introduced into the logistic regression model and corresponding results

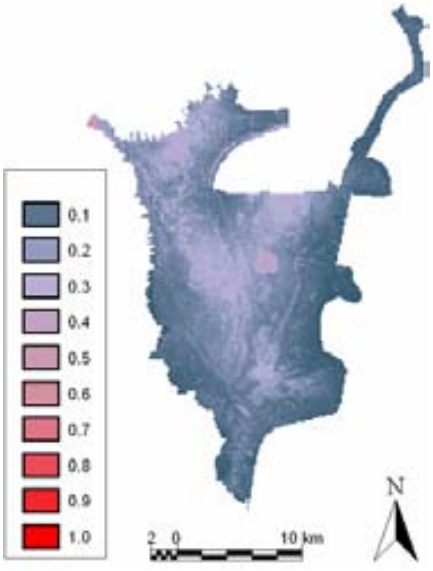


Figure 6: Logistic regression map for the probability of the presence of the salt-marsh bulrush (2007-2008).

3.2 Web portal

The maps and models generated have been integrated automatically into an existing Web Service: Servidor de Cartografía Digital de Seguimiento del Parque Nacional de Doñana (http://mercurio.ebd.csic.es/seguimiento). The existing interoperability protocols of the OGC (Open Geospatial Consortium [2008]) have been preserved, and a new tool specifically developed as a Web Processing Service (WPS) (Schut [2007]) has been added: analytic statistical overlay between a user's layer of polygon features and the map results generated in the present work. Figure 7 shows this portal.

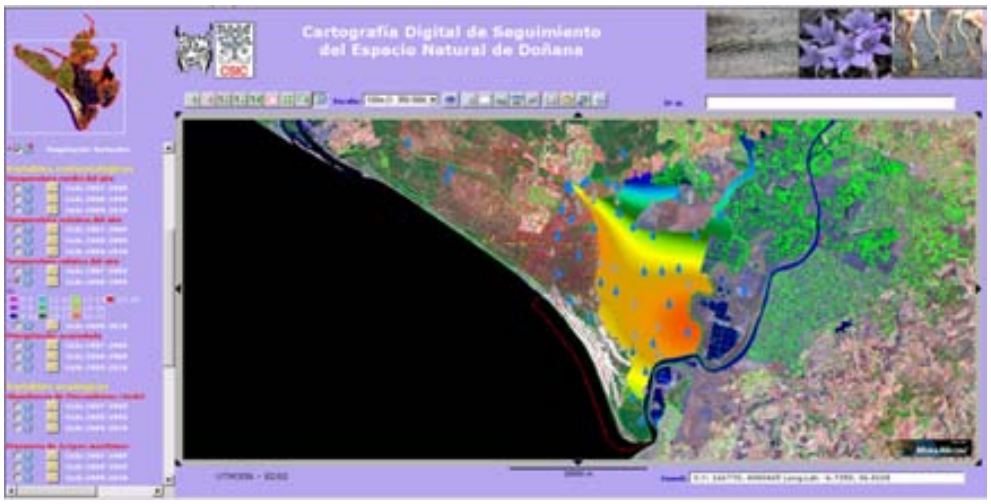


Figure 7. Web portal of the *Servidor de Cartografía Digital de Seguimiento del Parque Nacional de Doñana*.

4 CONCLUSIONS

The main contribution of this work is the automation and integration in a GIS environment of the entire proposed methodology, from downloading selected census data to the final publication of maps in a Web Map Server on the Internet using Open Geospatial Consortium standards. A service for invoking remote Web Processing Services for generating maps on demand is also provided.
The selected pilot cases, which represent different types of species census data and involve different estimation methods, demonstrate real implementations of the present work and a test-bed evaluation for extending this proposal to other scenarios.

ACKNOWLEDGMENTS

This work was supported by the Catalan Government (Spain) under grant SGR2009-1511 and the ICREA-Acadèmia award, by the Instituto de Cartografía de Andalucía (Spain) under grant "Ayudas a la investigación en materia de información geográfica" and uses data provided by the project CGL2009-09801 financed by FEDER funds.

REFERENCES

Bartier, P. M., Keller, C. P., Multivariate interpolation to incorporate thematic surface data using inverse distance weighting (IDW). *Computers & Geosciences*, 22(7): 795-799, 1996.

Bonham-Carter G.F, *Geographic information systems for geoscientists modelling with GIS*, Pergamon, 398 pp., 1994.

Burrough, P.A.,McDonnell,R.A., *Principles of Geographical Information Systems*. Oxford University Press, 333 pp, 1998.

Bustamante, J., Pacios, F., Díaz-Delgado R., Aragonés, D., Predictive models of turbidity and water depth in the Doñana marshes using Landsat TM and ETM+ images. *Journal of Environmental Management*. 90:2219-2225, 2009.

Díaz-Delgado, R., Bustamante, J., Aragonés, D. and Pacios, F., Determining water body characteristics of Doñana shallow marshes through remote sensing. In *Proceedings of the 2006 IEEE International Geoscience & Remote Sensing Symposium* (IGARSS2006), Denver, Colorado, EE.UU., 3662-3664. 2006.

Díaz-Delgado, R., An integrated monitoring programme for Doñana Natural Space: The set-up and implementation. In C. Hurford, M. Schneider e I. Cowx, (Ed.), *Conservation Monitoring in Freshwater Habitats: A Practical Guide and Case Studies.* Springer, Dordrecht, 375-386 pp., 2010.

Gutiérrez Martínez, J. M., Palacios, F. and Gutiérrez de Mesa, J.A., *El estándard XML y sus tecnologías asociadas*. Ed. Danysoft, pp 506, 2003.

Hancock, P.A., Hutchinson, M.F., Spatial interpolation of large climate data sets using bivariate thin plate smoothing splines. *Environmental Modelling and Software* 21, 1684-1694, 2006.

Hengl T. Finding the right pixel size. *Computers & Geosciences* 32:1283–1298, 2006.

Horning, N., Fosnight, E., eds. Secretariat of the Convention on Biological Diversity, Montreal,Technical Series no. 32, 201 pages. Pp. 83-102. ISBN: 92-9225-072-8

Isaaks, E.H., Srivastava, R.M. *Applied Geostatistics*. Oxford University Press, New York, 1989.

Kiehle, C., Greve, K., Heier, C. Requirements for next generation spatial data infrastructures-standardized web based geoprocessing and web service orchestration *Transactions in GIS* 11(6): 819-834, 2007

Kleinbaun, D.G, *Logistic regression*. New York, Springer-Verlag, 1994.

Lloyd, C. D., *Local Models for Spatial Analysis*. CRC Press, 244 pp, Belfast, 2006.

Masó J., Pons. X., Adding functionalities to WMS-WCS Clients: Download And Animation, *International Cartographic Conference*, A Coruña, 9-16, 2005.

Mitasova, H., Mitas, L., Interpolation by Regularized Spline with Tension. *Mathematical Geology*, 25 (6) 641-655 pp, 1993.

Microsoft Corporation Using BATCH files http://www.microsoft.com/resources/documentation/windows/xp/all/proddocs/en-us/batch.mspx?mfr=true [Date of access: 2-2-2012], 2010.

Ninyerola M, Pons X, Roure JM, A methodological approach of climatological modelling of air temperature and precipitation through GIS techniques. *International Journal of Climatology*, 20:1823-1841, 2000.

Open Geospatial Consortium: *OGC Reference Model*, Open Geospatial Consortium Inc. Reference number: OGC 08-062r4 Version: 2.0, 2008.

Pesquer L., Cortés A., Pons X., Parallel ordinary kriging interpolation incorporating automatic variogram fitting, *Computers & Geosciences* 37, 464-473, 2011.

Pons, X., *MiraMon. Geographical Information System and Remote Sensing Software*. Centre for Ecological Research and Forestry Applications, CREAF, ISBN: 84-931323-4-9 In Internet: http://www.creaf.uab.es/MiraMon, 2000.

Rouse J.W., Haas, R.H., Schell, J.A., Deering, D.W., Monitoring Vegetation Systems in the Great Plains with ERTS. *Third ERTS Symposium*, NASA SP-351, I:309-317, 1973.

Scotts, D., Drielsma, M. Developing landscape frameworks for regional conservation planning; an approach integrating fauna spatial distributions and ecological principles. *Pacific Conservation Biology* 8(4): 235-254, 2003

Valley, R.D., Drake, M.T., Anderson, C.S. Evaluation of alternative interpolation techniques for the mapping of remotely-sensed submersed vegetation abundance. *Aquatic Botany* 81(1):13-25, 2005

Walker Johnson, G., Gaylord, A.G., Franco, J.C., Cody, R.P., Brady, J.J., Manley, W., Dover, M., Garcia-Lavigne, D., Score, R., Tweedie, C.E., Development of the Arctic Research Mapping Application (ARMAP): Interoperability challenges and solutions *Computers & Geosciences* 37(11): 1735-1742, 2011.

Capítol 4

Spatial pattern alterations of JPEG2000 lossy compression in remote sensing images. Massive variogram analysis in High Performance Computing

Aquest capítol és una reproducció de Pesquer L., Pons X., Cortés A, Serral, I. (2013) “Spatial pattern alterations of JPEG2000 lossy compression in remote sensing images. Massive variogram analysis in High Performance Computing” *Journal of Applied Remote Sensing* 73595.

Spatial pattern alterations from JPEG2000 lossy compression of remote sensing images: massive variogram analysis in high performance computing

Lluís Pesquer,[a] Xavier Pons,[b] Ana Cortés,[c] and Ivette Serral[a]
[a]CREAF, Cerdanyola del Vallès 08193, Spain
l.pesquer@creaf.uab.cat
[b]Universitat Autònoma de Barcelona, Geography Department, Edifici B,. E-08193 Bellaterra, Catalonia, Spain
[c]Universitat Autònoma de Barcelona, Computer Architecture and Operating Systems Department, Edifici Q, E-08193 Bellaterra, Catalonia, Spain

Abstract. We evaluate the implications of JPEG2000 lossy compression of remote sensing images for spatial analytical purposes. The main issue is to identify which cases and conditions in geostatistical studies are suitable for using lossy compressed images. For these purposes, an extensive test using Landsat, compact airborne spectrographic imager (CASI), and moderate resolution imaging spectroradiometer (MODIS) image series has been analyzed, through applying and comparing two-dimensional and three-dimensional (spectral and time domains) compression methods with a wide range of compression ratios for several dates, different landscape regions, and spectral bands. Due to the massive test bed and consequently to the high time consuming executions, a parallel solution was specifically developed. Variogram analyses showed that all the compression ratios maintain the variogram shapes, but high compression ratios (>20:1) degrade the spatial patterns of the remote sensing images. These alterations are lower for the three-dimensional compression method, which was a considerable improvement (25%) on the two-dimensional method for large three-dimensional series (CASI, MODIS). However, the two methods behave similarly in the Landsat case. Finally, the parallel solution in a distributed environment demonstrates that high performance computing offers a suitable scientific platform for highly demanding time execution applications, such as geostatistical analyses of remote sensing images. [DOI: 10.1117/1.JRS.7.073595]

Keywords: geostatistics; remote sensing images; lossy compression; high performance computing.

Paper 12062P received Mar. 12, 2012; revised manuscript received Oct. 23, 2012; accepted for publication Dec. 14, 2012; published online Jan. 17, 2013.

1 Introduction and Objective

Remote sensing provides large amounts of data to the scientific community in terms of spatial, spectral, radiometric, and temporal resolutions[1] that can be used in a vast range of applications in agriculture,[2] climate change research,[3] urban management,[4] ecosystems monitoring,[5] and thematic mapping,[6] among others. It is a very interesting field for evaluating the quality of lossy compression techniques.

Lossy data compression facilitates accessing, sharing, and transmitting huge spatial datasets in environments with limited storage or with limited bandwidth. JPEG2000 is one of the possible lossy compression algorithms, and is currently an ISO standard.[7] However, lossy data compression procedures modify the original information,[8] and therefore rigorous studies are needed to understand the effects and consequences of this manipulation. Some previous works have

analyzed these alterations in remote sensing images in different fields and with varying objectives: spectral analysis,[9] digital classification,[10] texture analysis,[11] stereoscopy,[12] and multivariate regression,[13] among others.

The present work aims at studying the effect of JPEG2000 lossy compression by determining the differences between the spatial pattern domains in the original image and the compressed image, and more particularly, the impact on the geostatistical usage of remotely sensed imagery. Comparing the geostatistical properties of compressed images before and after compression is a different and novel approach, explored here instead of the usual comparison of the global parameter peak signal to noise ratio (PSNR).[14]

Exploring and describing the spatial variation in images is one of the main applications of geostatistics in remote sensing[15] as it can provide parameters for describing spatial patterns,[16] measurements for spatial autocorrelation,[17] procedures for downscaling images,[18] tools for estimating continuous variables,[19] data for radiometric coregionalization analysis,[20] and optimal sampling designs for ground surveys.[21] The variogram is an appropriate tool often used in geostatistics to carry out exploratory analyses.[22] Section 2.4 of the present work outlines this geostatistic tool and its specific characteristics when it is applied to remote sensing images. It is, however, a particularly slow procedure for processing large amounts of remotely sensed data. This computational constraint makes it necessary to use distributed environments, such as high performance computing (HPC). HPC provides methods and infrastructures for distributing computation in order to reduce the total execution time. Some examples of the benefits of HPC applied to remote sensing are hyperspectral image modeling,[23] fire monitoring,[24] meteorological applications,[25] image processing,[26] some classification methods[27] and web grid environments,[28] among others.

In this work, parallel computing plays an important role in allowing us to perform spatial pattern analyses within an acceptable period of time. Section 2.5 details the architecture, design paradigm and language, and programming libraries, as well as results in time execution performance in order to generate variogram analyses in an efficient scientific environment.

Many different cases were studied in order to obtain robust conclusions regarding the main questions raised: is the degradation similar at short and far distances applying different compression ratios? Do the analyzed compression methods modify patterns at different directions? Do possible pattern alterations depend on three or two dimension compression methods? Are the short and large remote imagery series behaviors similar in spectral and time dimension?

The rapid generation of variograms allowed applying these massive analyses to:

- Types of sensor images: multispectral series (Landsat), hyperspectral series (CASI), and large time series (MODIS) in a spatial resolution range from 3 to 250 m.
- Landscape regions with different spatial patterns.
- Scenes with different season phenology.
- A wide range of compression ratios.

This complete test bed is detailed in Sec. 2.1.

These different scenarios allow studying the possible spatial pattern alterations generated by the lossy compression methodologies and exploring the magnitudes and properties of these hypothetical changes. The conclusions of the work provide recommendations in each specific case about applying JPEG2000 lossy compression algorithms to remote sensing images for geostatistical analysis purposes.

2 Methodology

The proposed methodology is a chain of image processing, lossy compression procedures, and geostatistical analyses in an HPC environment. Figure 1 shows the different stages and elements of this methodology as well as the position of the parallel task within the processing chain. These stages are detailed in the following subsections.

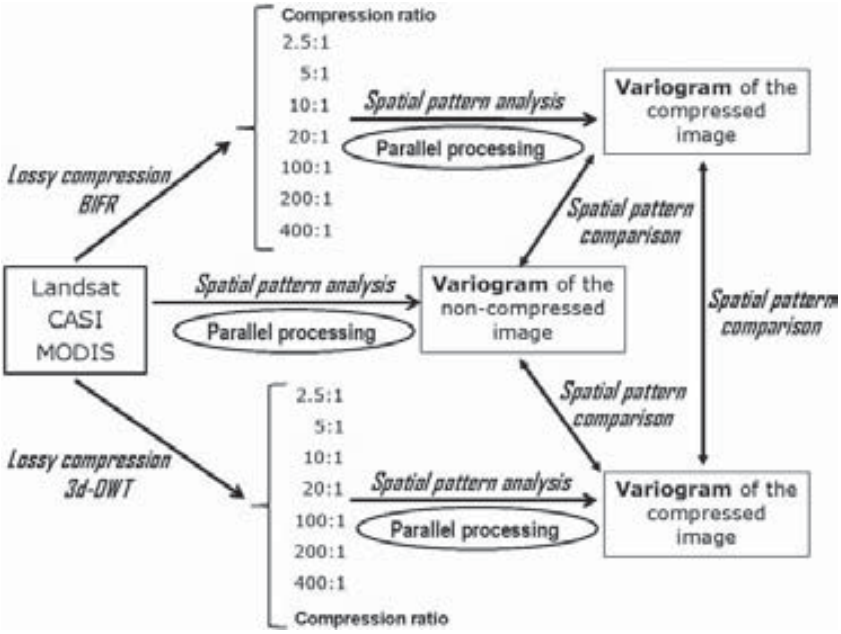


Fig. 1 Flowchart of the proposed methodology.

2.1 *Materials*

2.1.1 *Landsat*

In this study, four scenes of multispectral (seven bands) Landsat-5 Thematic Mapper (TM) images have been used.[29] The images selected for this work correspond to the dates 13-4-2006, 02-7-2006, 19-08-2006, and 11-9-2006, and cover two areas of about 15 km × 15 km (spatial resolution of 20 m) of the 197-031 and the 198-31 path-rows. They thus provide a set of images that is very representative of the main phenological cycle in this Mediterranean region. These two areas, in the Ripollès and Penedès counties, (Fig. 2, pink

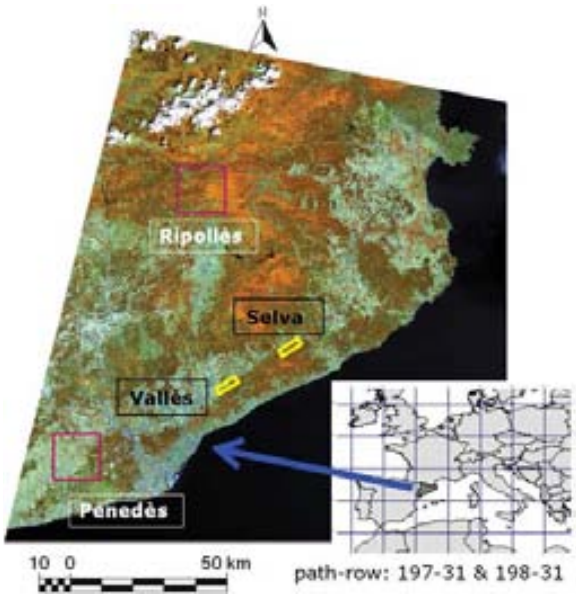


Fig. 2 Study regions: the Ripollès and Penedès areas for Landsat and MODIS are shown as pink rectangles and the Vallès and Selva areas for CASI images are shown as yellow rectangles.

squares) are located in Catalonia, a region of approximately $32,000$ km^2 situated in the northeast Iberian Peninsula at the extreme southwest of Europe (map location Fig. 2). The main landscape of the Ripollès region is deciduous forest, while the Penedès region has a vineyard landscape. Landsat images were chosen because they are the remote sensing images most used for research purposes in different applications,[30] and they are also used in a large variety of applied projects.[31]

2.1.2 *CASI*

The compact airborne spectrographic imager (CASI) is an optic sensor for hyperspectral scanning based on a small CCD bar. In this study, two images (provided by the Institut Cartogràfic de Catalunya, ICC) with 72 spectral bands from red to near infrared (409.15 to 955.70 nm and an average bandwidth of 7.6 nm) were used. [32] The two scenes selected correspond to 18-05-2007 and 19-06-2007. For these two dates, two different study areas of 4.3 km $\times$ 1.2 km at a spatial resolution of 3 m (Fig. 2, yellow rectangles) called Vallès and Selva located in the Montseny region, a mountain landscape in Catalonia. CASI images were chosen as an interesting test bed for three-dimensional compression methodologies (hyperspectral images are large radiometric third dimension examples).

2.1.3 *MODIS*

A moderate resolution imaging spectroradiometer (MODIS) Surface Reflectance Daily L2G Global 250 m product[33] was used in this study as an example of a large time series dataset used for analyzing the geostatistical patterns of compressed and noncompressed images. The time series include 73 selected cloud free images from 2007 (from 5-5-2007 to 30-09-2007) from the same regions as the two Landsat images analyzed (Penedès and Ripollès), but covering a more extensive area. Higher spatial resolution MODIS bands were used, i.e., the near infrared and red spectral bands. The region covered is a square of 25×25 km^2 centered in the 15×15 km^2 square of the corresponding Landsat dataset. The study area of MODIS images was expanded in order to obtain a number of pixels with a resolution of 250 m large enough for statistical purposes, in this case 100 columns $\times$ 100 rows.

In summary, this remote sensing material was selected because Landsat images are used most widely, CASI serves to extend the third spectral dimension, and MODIS serves to extend the third time dimension tests. Both CASI and MODIS provide large datasets, which improves the lossy compression analyses.

2.2 *Image Processing*

Different image processing methods were applied for each type of image depending essentially on the previous image processing level and on their particular characteristics.

Landsat images were acquired at the L1G processing level, and were then georeferenced and radiometrically corrected with MiraMon GIS software[34] according to the methodologies detailed in Refs. 35 and 36, and applied to the *SatCat* server.[37] The optical reflectance range of the corrected images extends from 0 to 10,000, representing the % of reflectance*100, and so numbers were written with two decimal places in a short integer. This factor and type of data (signed short integer) are also used to represent the possible range of ground temperatures in °C (derived from the radiometric correction) of the TM thermal band. All images use a -999 value to represent nodata regions, which in this case are normally caused by sensor problems or by post-processing artifacts. Defining the nodata value as -999 is not arbitrary because nodata values that are defined too far from the valid data range can negatively affect normalization compression procedures.[38] Consistent treatment of nodata values is one of the keys to performing correct image processing and subsequent data analysis.

However, CASI images were acquired geometrically georeferenced using the SISA system developed by ICC.[39] The flight orientation was SW-NE, which resulted in the geometrically corrected image being very large ($18,366$ columns $\times$ $11,918$ rows), and therefore the CASI images were rotated -32.5 degrees with respect to the north projection in order to reduce the large amount of nodata pixels. The value range, data type and nodata definition in the

CASI images were unified to match those of the Landsat images, also using the MiraMon software.

Finally, the MODIS images were downloaded from the Warehouse Inventory Search Tool[40] as a georeferenced reflectance product. It was only necessary to clip them to the study regions already detailed in the previous section and unify the nodata values.

2.3 *Lossy Compression*

The reflectance product obtained was compressed at a wide range of different quality levels with the following compression ratios (CR): 2.5:1, 5:1, 10:1, 50:1, 100:1, 200:1, and 400:1 (from soft to hard compression). Two methodologies, based on the JPEG2000 lossy compression procedures, were evaluated in this work:

- The band-independent fixed-rate (BIFR) method is an independent compressor in which no inter-band redundancy is exploited and the bit-rate is split equally among all bands. BIFR is therefore a two-dimensional (space) compression method.
- Three-dimensional discrete wavelet transform (3d-DWT) is an inter-band decorrelation technique that exploits the spectral (or temporal) redundancy between multiple bands or scenes: the third dimension. This transform is applied to the input image to obtain a spectrally (or temporally) decorrelated image.

Both compression procedures were applied using the Kakadu software,[41] which is a complete implementation of the JPEG2000 standard, Part 1, completed with a great deal of Part 2 and 3.[42] The BOÍ software,[43] an implementation of the JPEG 2000 (Part 1) standard, was used for calculating the PSNR quality compression indicator in reference to noncompressed images.

2.4 *Geostatistics*

Geostatistics is a subset of statistics specialized in the analysis, interpretation, and inference of geographically referenced data.[44] It was initially developed by Matheron[45] and defined as a set of methods for studying the spatial distribution of a variable from a statistical approach in order to estimate the corresponding unknown value of a particular location or simulate its variability. When this variation has a spatial structure, the regionalized variable theory makes it possible to take spatial properties into account following a stochastic approach, and assumes a constant local mean and a stationary variance of the differences between regions separated by a given distance and direction.[46] The variogram (also referred to as semivariogram) is based on these previous assumptions and, as defined in Eq. (1), it plots the variance function depending on sample distance separation and, consequently, the spatial patterns analyzed. This expression implies, for remote sensing images, that pixel values variance is only a function of distance and direction between pixels; then, there is no dependence on their particularly locations and, in statistics sense, this variance pattern is stationary for all study regions. Therefore, the variogram describes the pattern distribution of image spatial resolution, and it can identify a study region, a subscene of remote sensing image in this work.

$$\gamma(h) = \frac{1}{2 \cdot n} \sum_{i=1}^{n} [z(\vec{x}_i) - z(\vec{x}_i + h)]^2. \tag{1}$$

Equation 1: variogram definition. Note that the dependence is on distance (h) and on the squared difference of z values (pixel values), and independent of particular positions (x_i and $x_i + h$).

The sampled variogram can be modeled and fitted by a continuous function to identify structural parameters that characterize the spatial pattern for any quantitative variable distribution, including, of course, those of remotely sensed images, as carried out here in this study and in other previous studies[47] (Fig. 3). These parameters are:

- Nugget: the variance near the origin, so at very low distances, it represents the fluctuations at scales smaller than the sampling interval (lag distance), and the component of the nonspatially correlated error.

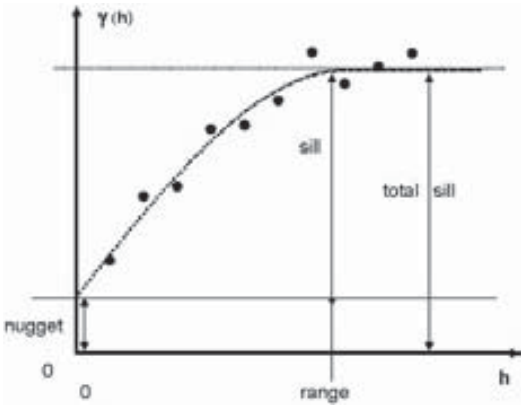


Fig. 3 Variogram structure parameters.

- Range: the distance at which the variogram reaches saturation. It means that variance becomes stabilized despite distances increase and it defines the limit distance of autocorrelation.
- Sill (partial): the variance at the variogram saturation excluding nugget variance, it means a measure for spatially correlated variability.
- Total sill: sum of the partial sill and nugget. It represents the global (random + spatially correlated) variability at the autocorrelation threshold.

For each specific geostatistical application, the key role parameter depends upon the focus of a particular spatial pattern analysis. For example, range is the most suitable for autocorrelation analysis, nugget for determining stochastic noise and sill for variability's measurements.

The total sill was chosen as the main quality measure shown in Sec. 3 for analyzing alterations in the spatial patterns at studied compression ratios and with different methodologies. Different selected plots were used for three sensors for several dates, bands, and regions, which made it possible to determine the complete variogram structure by comparing the theoretical plot in Fig. 3 with the experimental plots in Sec. 3.

2.5 *Parallel Solution*

Most variogram analyses are used to explore the spatial patterns of irregularly distributed samples.[48] Variogram analysis is usually a previous step to interpolation (kriging), which is a process with a high computational cost.[49]

Nevertheless, the variogram analysis of remotely sensed images has higher computing demands to those of irregularly distributed samples. Indeed, a large number of pixels are involved even in geographically small scenes with a high or medium spatial resolution, and therefore a very large number of pairs are necessary to obtain the variogram. Table 1 compares four examples of the amount of data involved and the time necessary for executions running on a single, normal PC (Intel(R) Xeon (R) 3.0 GHz and 1 GB of RAM).

In this work, a large number of spatial pattern analyses have been executed. All spectral bands (the number of bands for Landsat, CASI, and MODIS are 7, 72, and 2, respectively) on several scene dates (4, 2, and 73) and in two different study regions were analyzed. In addition, the images were compressed at several compression ratios (eight in all cases) and compared with noncompressed images. This implied 504 analyses for Landsat, 2592 for CASI, and 2628 for MODIS. Performing all these analyses would be a very time-consuming process; a single CASI analysis takes 57,727 s (see Table 1). Therefore, a distributed environment was the most suitable solution for a complete and efficient study.

In order to reduce the execution time, the authors carried out a parallel implementation of the variogram analyses. The master/worker programming paradigm was used in which the master process schedules, distributes, and coordinates the tasks executed by the worker processes. The code was written in ANSI-C language, including MPI (message passing interface) as a message-passing library (version 1.2.7) that is used by the master process to communicate with the worker

Table 1 Comparison of processing time of variogram analyses between two irregularly distributed point measurements and remote sensing images. The sparse irregular distribution corresponds to a network of weather stations, and the dense, irregular, distribution to a GPS network of altimetry measurements, both detailed in Ref. 44; the remotely sensed image corresponds to two (Landsat and CASI) of the three examples used in this work.

Type	Extension (km^2)	*N* data	*N*. pairs examined	Time (s)
Sparse point irregular sampling	69,284.96	100	1070	26
Dense point irregular sampling	0.216	2855	1,627,350	946
Remotely sensed image: medium resolution	225	562,500	16,493,978,252	43,850
Remotely sensed image: high resolution	6.5	725,145	33,554,502,751	57,727

processes. Since MPI has become a standard de-facto message passing library,[50] this solution (ANSI-C + MPI) guarantees portability to different computer platforms.

The distributed load design[51] is not specifically defined and optimized for the characteristics of the images used in this work. However, it is a flexible design, tested with a wide range of image dimension that maintains a good balance and scalability suitable for other studies with a wide range of image dimensions and cluster properties. The scalability solution (adapting point samples analysis from Ref. 49 to remote sensing image analysis in the present work) can be achieved by defining a relatively small load unit: of two image rows of pixels. Each worker analyzes all possible combinations of nonrepeated pairs between two rows of the image and returns the partial variance and distance of the pairs involved to the master. The master distributes tasks (rows of data) to workers, accumulates partial results and manages input (remotely sensed image) and output (resulting variogram) tasks.

3 Results

Two types of results are presented in this section: geostatistical and computational. The geostatistical results are a set of variograms that compare images compressed at different ratios and noncompressed images for three image types (Landsat, CASI, and MODIS) focusing on the spectral or time dimension and possible anisotropy. The total sill parameters of the theoretical variogram are shown in different tables, while the experimental variograms are shown in plots.

The computational results are focused on performance analyses of the parallel solution. The example shown in Sec. 3.2 corresponds to Landsat executions, but the performance results are very similar for all the image types analyzed.

3.1 *Variogram Analysis*

As explained above, lossy compression mechanisms modify the original data information. Figure 4 illustrates the effects of compression ratios on quality visualization of the images. In this figure, the central image is a noncompressed CASI example, the top image sequence corresponds to three BIFR compressed images; from left to right, it is shown increasing compression ratios and decreasing quality. The bottom image sequence corresponds to the same sequence improved using 3d-DWT techniques. This composition summarizes most of the results presented in this section, and the following tables and figures quantify these effects in a geostatistical sense using the variogram spatial pattern analysis.

3.1.1 *Landsat*

The first plot of Fig. 5 is a representative result of the alterations in the spatial pattern caused by the lossy compression (BIFR in this case) mechanisms. The main alteration to the spatial pattern due to lossy compression is the reduction in data variability at high compression ratios (more than 100:1), although at lower CRs (from 2.5:1 to 50:1) the variograms have very similar

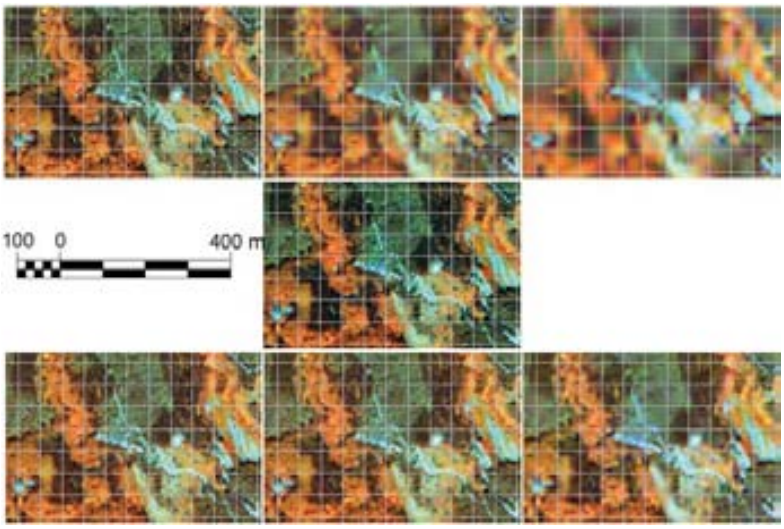


Fig. 4 The effects of compression on a CASI image fragment from the Selva region: noncompressed in the middle, upper sequence of BIFR compression ratios: 2.5:1, 10:1, and 400:1. The images in the bottom sequence were compressed at the same ratios but the 3d-DWT method was used. All of them correspond to a 4-5-3 band composite image.

shapes and structural parameters (nugget, range and sill). Table 2 is a summary of the behavior of the total sill, which is the most representative parameter in this study, for two dates and two regions for the BIFR method. Comparing the sill between compressed and noncompressed images represents a measurement of the loss of detail and variability caused by a compression method for all Landsat spectral bands. Table 3 shows the same information in relation to the 3d-DWT compression method. As seen in Fig. 6, there are only small differences between the two methods for Landsat images with the highest CR. Therefore, the following tables and figures corresponding to Landsat images only show the results for the BIFR method. The next cases (CASI and MODIS) demonstrate that 3d-DWT is a more useful method for image series that are larger than Landsat image series.

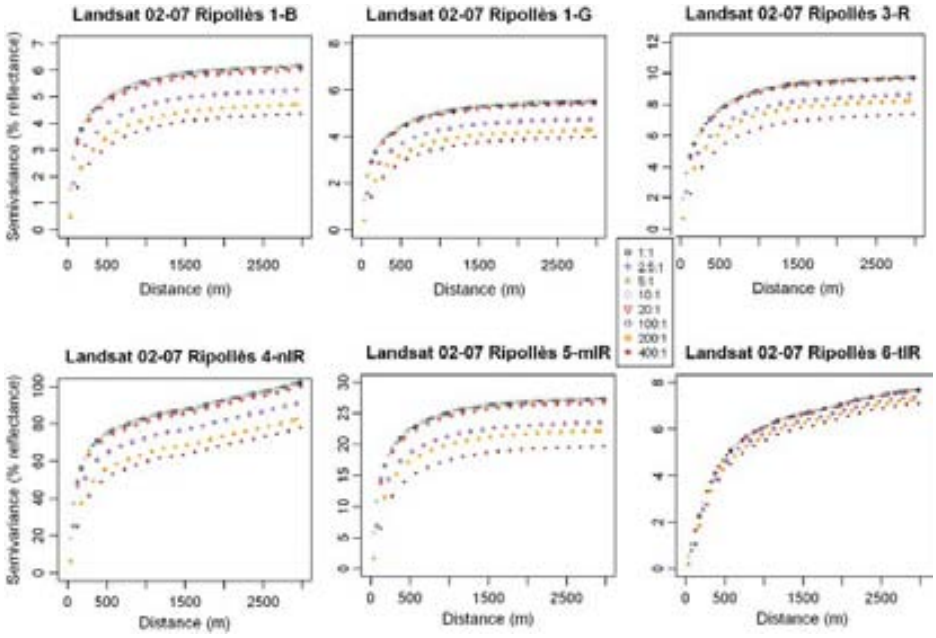


Fig. 5 Variogram results of the BIFR compression method on a Landsat image from the Ripollès region on 02-07-2006, a multispectral view.

Table 2 Two examples (Ripollès region on 02-07-2006, and Penedès region on 11-9-2006) of the variation in the variogram parameter total sill at different compression ratios for the BIFR method.

	Band/CR	1:1	2.5:1	5:1	10:1	20:1	100:1	200:1	400:1
Ripollès02-07-2006	**1-B**	5.92	5.95	5.96	5.95	5.80	5.04	4.51	4.17
	2-G	5.30	5.34	5.35	5.33	5.23	4.58	4.13	3.81
	3-R	9.40	9.44	9.44	9.43	9.31	8.3	7.90	7.07
	4-nIR	92.96	93.30	93.34	93.19	90.93	82.33	74.19	69.57
	5-mIR1	26.31	26.40	26.43	26.38	25.77	22.63	21.28	18.85
	6-tIR	7.10	7.10	7.11	7.11	7.09	6.84	6.72	6.53
	7-mIR2	15.74	15.78	15.81	15.69	15.40	13.8	13.41	12.50
Penedès11-9-2006	**1-B**	18.01	18.51	18.57	18.53	18.06	15.68	15.68	15.68
	2-G	17.30	17.73	17.79	17.77	17.33	14.96	14.96	14.95
	3-R	25.94	26.37	26.43	26.31	25.68	22.21	22.21	22.21
	4-nIR	36.49	37.8	38.00	37.49	36.23	30.33	30.33	30.32
	5-mIR1	44.61	45.53	45.67	45.11	43.76	37.43	37.43	37.43
	6-tIR	3.53	3.54	3.55	3.53	3.53	3.35	3.35	3.36
	7-mIR2	31.54	32.06	32.17	31.9	31.15	27.24	27.23	27.24

Table 3 Two examples (Ripollès region on 02-07-2006, and Penedès region on 11-9-2006) of the variation in the variogram parameter total sill at different compression ratios for the 3d-DWT method.

	Band/CR	1:1	2.5:1	5:1	10:1	20:1	100:1	200:1	400:1
Ripollès02-07-2006	**1-B**	5.92	5.95	5.97	5.93	5.88	5.54	5.35	5.02
	2-G	5.30	5.33	5.34	5.33	5.30	5.05	4.89	4.61
	3-R	9.40	9.42	9.45	9.45	9.37	9.06	8.87	8.65
	4-nIR	92.96	92.80	92.83	92.19	90.44	82.75	76.49	70.17
	5-mIR1	26.31	26.37	26.40	26.04	25.33	22.47	20.42	18.34
	6-tIR	7.10	7.08	7.10	7.20	7.29	7.31	7.28	7.16
	7-mIR2	15.74	15.76	15.80	15.80	15.62	14.37	13.63	12.84
Penedès11-9-2006	**1-B**	18.01	18.45	18.52	18.59	18.38	16.81	16.37	15.69
	2-G	17.30	17.64	17.72	17.71	17.53	16.33	15.76	14.93
	3-R	25.94	26.26	26.35	26.06	25.34	22.47	20.73	19.15
	4-nIR	36.49	37.62	37.77	37.40	35.99	31.00	27.46	23.84
	5-mIR1	44.61	45.33	45.43	44.98	43.53	39.05	36.35	32.06
	6-tIR	3.53	3.52	3.56	3.82	4.28	5.18	5.73	6.44
	7-mIR2	31.54	31.93	32.03	32.34	32.33	30.53	29.44	27.97

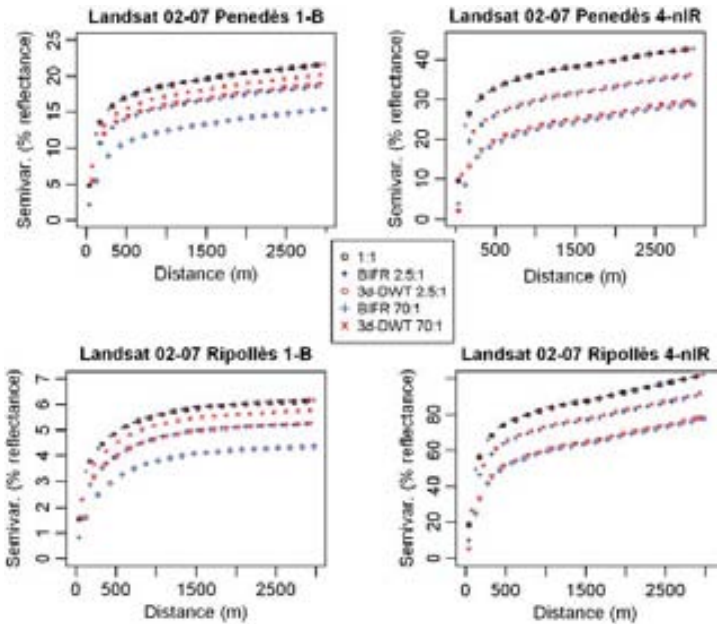


Fig. 6 Landsat variogram comparison between BIFR and 3d-DWT compression in relation to the original pattern.

The sequenced plot in Fig. 5 shows that there are no significant differences in the variogram shape between bands, although there are differences in band range variances, and consequently in structural parameters depending on the region of the landscape.

Table 4 and Fig. 7 reproduce the same variogram spatial pattern in a temporal sequence for an example of the near infrared (4-nIR) band.

Furthermore, lossy compression may produce different variance alterations in relation to the spatial direction. In order to analyze this possible anisotropy modification, variance pattern at different azimuth angles has been studied. Figure 8 shows that the lossy compression mechanisms used in this work maintain the expected directional patterns at all compression ratios; thus, this result proves that the previous variograms were always omnidirectional.

Table 4 Time comparison of variogram results (total sill) for BIFR compression of Landsat images from the Penedès region on the 1-B and nIR bands.

	Date	Band/CR	1:1	2.5:1	5:1	10:1	20:1	100:1	200:1	400:1
Penedès	**13/04/2006**	**1-B**	22.32	22.01	22.03	22.21	21.96	20.07	19.12	17.72
		4-nIR	38.46	38.48	38.48	38.23	36.88	31.36	28.10	24.63
	02/07/2006	**1-B**	20.51	20.12	20.12	20.29	20.21	18.68	18.37	17.42
		4-nIR	40.49	39.70	39.70	39.38	38.09	33.33	30.04	26.95
	19/08/2006	**1-B**	37.96	37.27	37.29	37.34	36.94	33.89	32.60	30.76
		4-nIR	54.20	53.35	53.38	53.18	51.74	45.71	42.01	37.39
	11/09/2006	**1-B**	18.01	18.45	18.52	18.59	18.38	16.81	16.37	15.69
		4-nIR	36.49	37.62	37.77	37.40	35.99	31.00	27.46	23.84

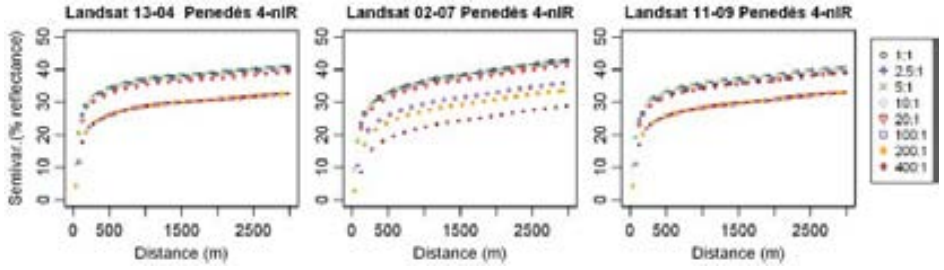


Fig. 7 Date comparison of variogram results for BIFR compression of Landsat images from the Penedès region on the near infrared (4-nIR) band.

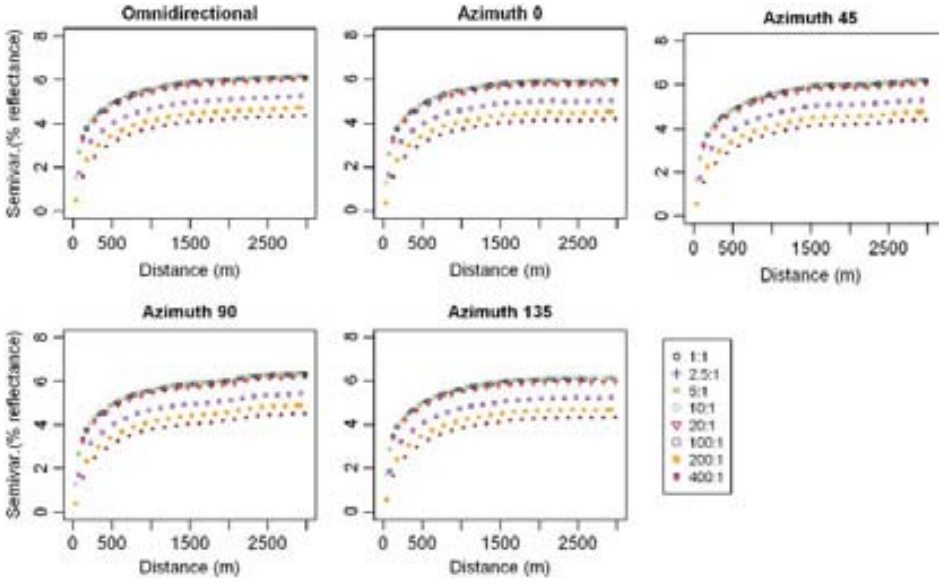


Fig. 8 Variogram behavior in relation to different azimuth directions. Landsat 02-07-2006 1-B band images.

3.1.2 *CASI*

The main characteristic of CASI images in relation to Landsat images is their higher spectral resolution (much more data in the spectral domain). The higher spatial and radiometric resolution are also notable, but they are not the focus of the present study. It is therefore an interesting test bed for exploiting three-dimensional compression methods. As it is impossible to show the results for 72 bands for two regions and two dates, the following figures display representative results for selected bands for different regions and dates. For example, Table 5 shows the total sill

Table 5 Total sill parameters for the k4p scene in the Selva region and for the complete sequence of compression ratios at four representative spectral bands.

	Band/CR	1:1	2.5:1	5:1	10:1	20:1	100:1	200:1	400:1
SelvaK4p	**10**	14.07	14.07	14.07	13.82	13.91	12.25	10.37	8.71
	21	36.88	36.39	36.41	36.21	35.13	29.81	26.40	23.62
	34	40.92	41.18	41.20	40.97	40.26	37.71	35.80	31.88
	58	398.05	398.05	397.98	395.84	388.54	370.07	356.27	334.35

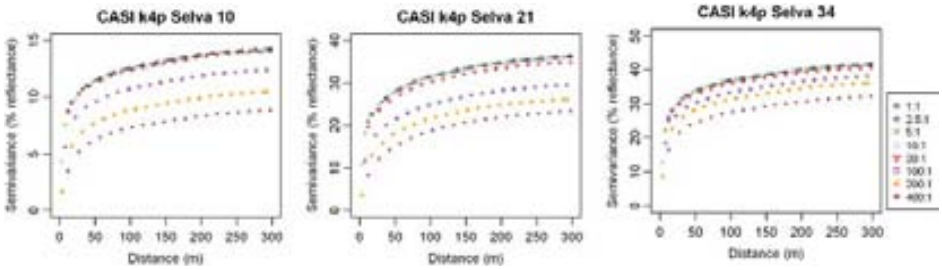


Fig. 9 Variogram results for the 3d-DWT compression method in the Selva region at selected CASI bands on 19-06-2007 (*k*4*p* scene).

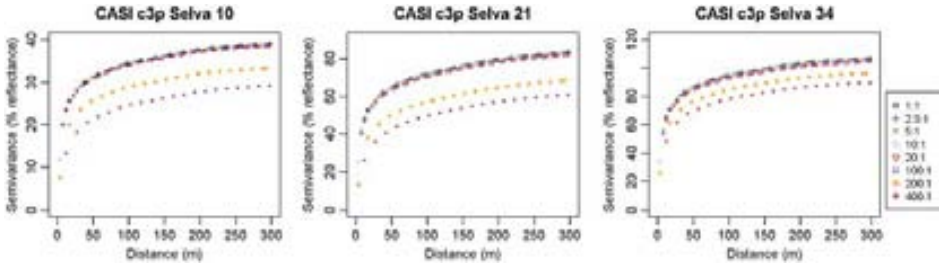


Fig. 10 Comparison of variogram results for the 3d-DWT compression method in the Selva region at selected CASI bands on 18-05-2007 (*c*3*p* scene).

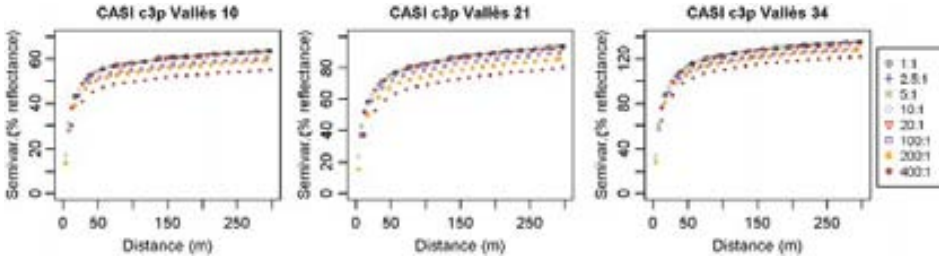


Fig. 11 Variogram results for the 3d-DWT compression method in the Vallès region at selected CASI bands.

parameter of the 3d-DWT method for four selected bands (10, 21, 34, and 58) in the *k*4*p* scene (19-06-2007), while Fig. 9 plots variograms at different compression ratios for three selected bands. Figures 10 and 11 show the variogram analyses of the *c*3*p* scene for the two regions applying the 3d-DWT compression method.

Table 6 shows the total sill parameter for the other scene (*c*3*p* on 18-05-2007) in the two study regions (Selva and Vallès) comparing the BIFR and 3d-DWT compression methods at selected CRs and bands (10, 21, 34). This table comparison demonstrates that spatial patterns have better fidelity with the 3d-DWT method than the BIFR method for CASI images. Figure 12 shows that 3d-DWT plots are closer to the noncompressed patterns, and that BIFR significantly alters the spatial patterns, especially for high compression ratios.

3.1.3 *MODIS*

The MODIS test bed allows a much larger amount of data to be explored in the time domain than the Landsat images because the MODIS daily revisit period makes it possible to obtain a large cloud-free time series. These series are especially appropriate for analyzing the pattern alterations of three-dimensional compression methods *versus* two-dimensional ones.

Table 6 Comparison between total sill parameters for the BIFR and 3d-DWT methods applied to the *c3p* scene in two regions at four representative compression ratios and three spectral bands.

			2.5:1		5:1		20:1		100:1	
c3p	Band/CR	1:1	BIFR	3d	BIFR	3d	BIFR	3d	BIFR	3d
Selva	**10**	38.65	38.65	38.66	38.69	38.67	37.24	38.31	30.59	38.31
	21	84.98	84.99	84.98	78.85	85.01	78.85	83.59	65.12	83.59
	34	105.63	105.97	105.93	101.46	106.01	101.46	103.88	85.42	103.88
Vallès	**10**	65.88	65.88	65.87	66.05	65.84	63.90	65.61	55.97	62.49
	21	99.38	99.38	99.38	99.40	99.36	98.46	98.46	82.55	95.26
	34	139.98	142.59	139.99	143.54	139.99	139.11	138.73	124.53	134.03

The results for the MODIS images are similar to those obtained for CASI images, but both are different from the Landsat images. In these cases, the 3d-DWT method improves the quality of BIFR compressions. Table 7 shows that, at low compression ratios, 3d-DWT and BIFR maintain the spatial variability patterns of the original images, and that, at high compression ratios, BIFR loses quality compared to the 3d-DWT method. The plots in Figs. 13 and 14 (corresponding to the same region but to different spectral bands) show a reduction in the total sill parameter at increasing compression ratios for the 3d-DWT method, while Fig. 15 shows that 3d-DWT produces the pattern variability more faithfully than BIFR.

In summary, these results quantify spatial pattern alterations for lossy compression images depending on compression ratios, compression methods, and series dimension through analyzing differences between structural variogram parameters (focusing on total sill) of compressed and noncompressed images. These results shows that, for compression ratios higher than 1:100, the variogram is clearly degraded, reducing variability and thus, some capabilities or accuracy in related applications. This degradation could be partially solved, in some cases, by 3d-DWT, but this method needs a large amount of redundant data for significantly improve the BIFR method. These improvements are similar for time series and spectral series.

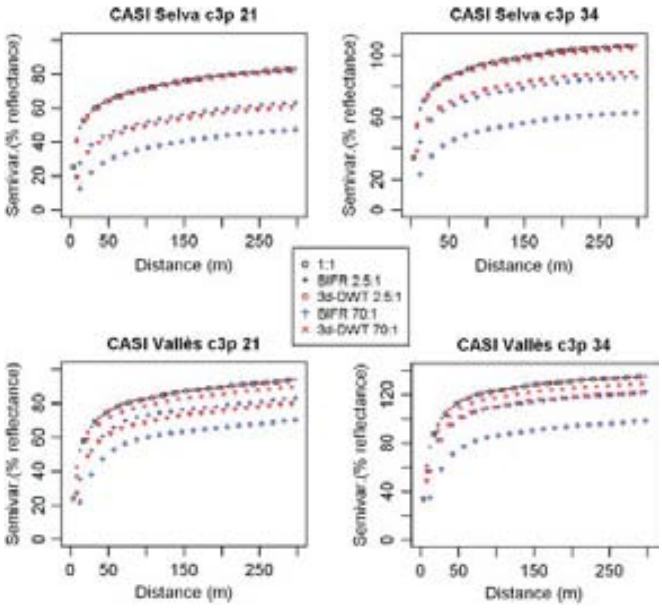


Fig. 12 Comparison of variogram results of the 3d-DWT and BIFR compression methods in both regions at two selected CASI bands.

Table 7 Comparison between the BIFR and 3d-DWT methods of the total sill for four selected dates in the Penedès region for two spectral bands at representative compression ratios.

Penedès	CR	1:1	2.5:1		20:1		70:1	
Date	Band		BIFR	3d	BIFR	3d	BIFR	3d
176	**RED**	20.04	20.06	20.02	19.10	19.10	12.50	19.45
	NIR	19.46	19.43	19.46	17.62	18.70	13.93	18.73
207	**RED**	20.23	20.27	20.21	20.20	20.20	11.67	19.70
	NIR	20.41	20.48	20.46	19.35	20.04	13.63	18.56
249	**RED**	19.77	19.86	19.92	19.90	19.90	11.05	18.52
	NIR	29.05	29.06	29.13	26.45	28.49	23.62	27.20
271	**RED**	16.43	16.42	16.43	15.83	15.83	10.13	13.97
	NIR	18.72	18.70	18.70	17.47	17.63	9.91	16.25

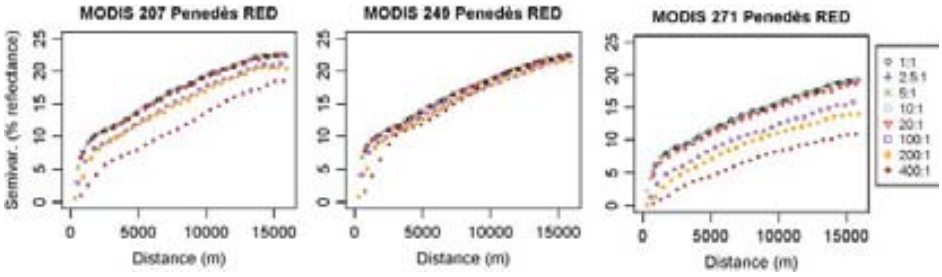


Fig. 13 Comparison of experimental variograms of the 3d-DWT compression method for the Penedès region on three selected dates applied to MODIS images in the red band.

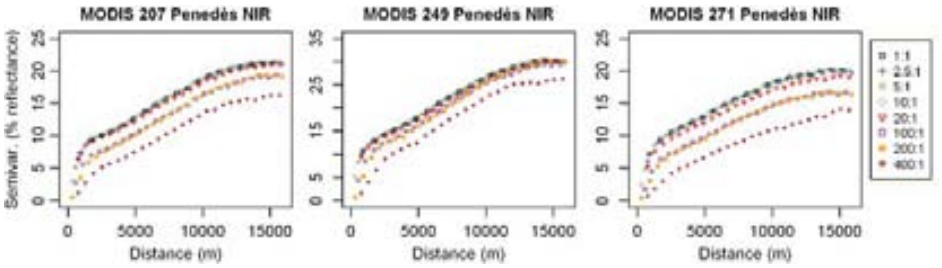


Fig. 14 Variograms of the 3d-DWT compression method for the Penedès region on three selected dates for MODIS images in the near infrared band.

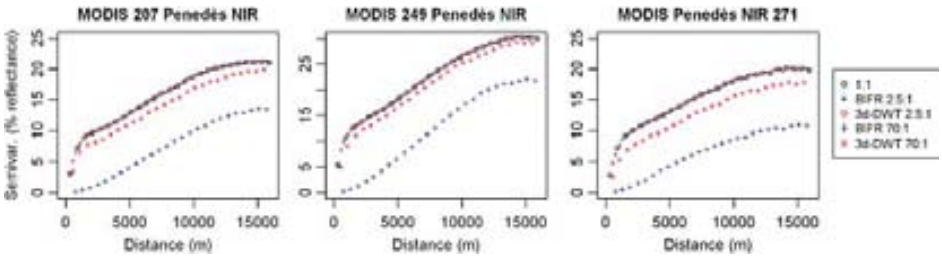


Fig. 15 Comparison of variograms for the BIFR and 3d-DWT methods for three selected dates in the Penedès region at representative compression ratios for MODIS images in the near infrared band.

This significant alteration (CR higher than 1 : 100) could be relevant in several geostatistical studies of remote sensing images, such as geostatistical applications referred in Sec. 1 of this work, and for example, in landscape scale classification based on variogram analysis.[52] In this example, if someone uses compressed remote sensing images instead of original images, and this compression is not done with the appropriate parameters, it could significantly alter classification results and, consequently, its main goal of determining the optimal support size (i.e., optimal spatial resolution) for characterizing forest ecosystems.

3.2 *Performance Results*

All the variogram analyses were processed using the IBM cluster of the research laboratory of the Computer Architecture and Operating Systems Department at the Universitat Autònoma de Barcelona. The IBM cluster is formed by 32 nodes, each with two Dual Core Intel(R) Xeon (R) 3.0 GHz processors with 12 Gbyte of RAM, and communicated with an integrated dual gigabit ethernet. Table 8 shows the average time of 14 independent executions using a different number of workers, while Fig. 16 shows the graphical representation and corresponding speedup evaluation.[53]

The proximity between the empirical speedup behavior and the theoretical linear speedup confirms that the parallel design and implementation provide a satisfactory solution for the computational problem at hand. These performance results demonstrate the validity of the proposed parallel solution, and the significant time reduction evidences the benefits of using distributed environments for processing large amounts of data, such as remotely sensed images. In fact, they provide an exhaustive test bed allowing running many executions with different parameters at

Table 8 Execution time and speedup using n (0 to 24) workers.

No. workers	Time (s)	Speedup
0	2975.03	
2	2112.33	1.41
4	1067.45	2.79
8	534.18	5.57
12	357.54	8.32
16	269.00	11.06
20	216.24	13.76
24	186.31	15.97

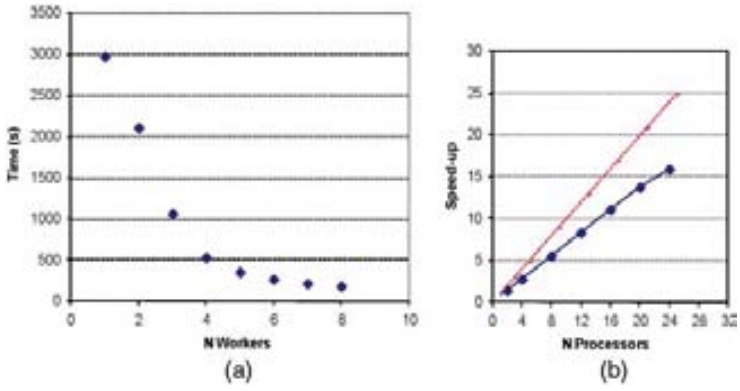


Fig. 16 Execution times (a) and speedup (b) with a different number of workers. The red dotted line corresponds to the theoretical linear speedup, while the blue line and points correspond to the empirical data.

several compression ratios in different regions on different dates, etc., and obtaining faster results and more agile comparisons.

4 Conclusions

The main conclusions of this study on the alterations in the spatial pattern produced by JPEG2000 compression methodologies applied to remote sensing images are drawn hereafter:

- Low compression ratios (less than 20:1 in the present study) maintain the radiometric image variability in all distance analyses, and therefore generate very similar variograms. This general behavior is slightly different for the 3d-DWT and BIFR methods: if the third dimension is large enough (CASI and MODIS) 3d-DWT is slightly more accurate to the noncompressed images than BIFR (less than a 0.5% reduction in variability).
- High compression ratios (over 20:1 in the present study) alter all spatial patterns of remote sensing images, but 3d-DWT is considerably better than BIFR for large three-dimensional images; 3d-DWT maintains about 25% more radiometric variability than BIFR.
- For Landsat images, seven spectral bands as a third spectral dimension are not enough to exploit three-dimensional compression, and therefore the BIFR method is preferable due to its simplicity.
- Despite the fact that pattern variability depends on the spectral band, all variogram alterations are very similar in all spectral bands and spatial directions.

Regarding to the computational methods developed in this work to reduce the execution time required for variogram analyses of remote sensing images:

- A distributed parallel solution based on a master/worker scheme and using MPI as the message-passing library obtains an efficient computing performance and provides a suitable environment for carrying out exhaustive analyses with very different compression ratios.

Acknowledgments

This work was partially supported by the Spanish government, by FEDER, and by the Catalan government, under Grants TIN2009-14426-C02-02, TIN2007-64974, PTA2010-3619-I, and SGR2009-1511. The authors are especially grateful to the Institut Cartogràfic de Catalunya, ICC, for providing the CASI data. Xavier Pons is a recipient of an ICREA Acadèmia Excellence in Research grant (2011 to 2015). Finally, the authors also wish to thank to the anonymous reviewers for their constructive comments that helped to improve the manuscript.

References

1. J. R. Jensen, *Introductory Digital Image Processing: A Remote Sensing Perspective*, 3rd ed., Prentice Hall Series in Geographic Information Science (2005).
2. P. Serra and X. Pons, "Monitoring farmers decisions on Mediterranean irrigated crops using satellite image time series," *Int. J. Rem. Sens.* **29**(8), 2293–2316 (2008), http://dx.doi.org/10.1080/01431160701408444.
3. P. Schneider and S. J. Hook, "Space observations of inland water bodies show rapid surface warming since 1985," *Geophys. Res. Lett.* **37**(22), L22405 (2010), http://dx.doi.org/10.1029/2010GL045059.
4. N. B. Chang et al., "Change detection of land use and land cover in an urban region with SPOT-5 images and partial Lanczos extreme learning machine," *J. Appl. Remote Sensing* **4**(1), 043551 (2010), http://dx.doi.org/10.1117/1.3518096.
5. K. Richter et al., "Plant growth monitoring and potential drought risk assessment by means of Earth observation data," *Int. J. Rem. Sens.* **29**(17), 4943–4960 (2008), http://dx.doi.org/10.1080/01431160802036268.

6. T. M. Lillesand, R. W. Kiefer, and J. W. Chipman, *Remote Sensing and Image Interpretation*, 5th ed., John Wiley and Sons, New York (2004).
7. D. S. Taubman and M. W. Marcellin, *JPEG 2000: Image Compression Fundamentals, Standards, and Practice*, pp. 778, Kluwer Academic Publishers, Dordrecht (2002).
8. T. Ranchin and L. Wald, "The wavelet transform for the analysis of remotely sensed images," *Int. J. Rem. Sens.* **14**(3), 615–619 (1993), http://dx.doi.org/10.1080/01431169308904362.
9. G. Martín et al., "Impact of JPEG2000 compression on endmember extraction and unmixing of remotely sensed hyperspectral data," *J. Appl. Rem. Sens.* **4**(1) 041796 (2010), http://dx.doi.org/10.1117/1.3474975.
10. A. Zabala and X. Pons, "Effects of lossy compression on remote sensing image classification of forest areas," *Int. J. Appl. Earth Observat. Geoinf.* **13**(1), 43–51 (2011), http://dx.doi.org/10.1016/j.jag.2010.06.005.
11. M. A. Chaudhry et al., "Optimization of wavelet bases for texture analysis," in *Proc. Fifth IEEE International Symposium on Signal Processing and Information Technology*, Athens, pp. 789–793 (2005).
12. Y. Li et al., "Adaptive compression of remote sensing stereo image pairs," *J. Appl. Rem. Sens.* **4**(1), 041777 (2010), http://dx.doi.org/10.1117/1.3495716.
13. A. Zabala et al., "Implications of JPEG2000 lossy compression on multiple regression modeling," *Proc. SPIE* **6749**, 674918 (2007), http://dx.doi.org/10.1117/12.738028.
14. M. Kiefner and M. Hahn, "Image compression versus matching accuracy," in *International Archives of Photogrammetry and Remote Sensing*, Vol. XXXIII, Part B2, pp. 316–323 (2000).
15. P. J. Curran and P. M. Atkinson, "Geostatistics and remote sensing," *Prog. Phys. Geogr.* **22**(1), 61–78 (1998), http://dx.doi.org/10.1177/030913339802200103.
16. C. S. A. Wallace, J. M. Watts, and S. R. Yool, "Characterizing the spatial structure of vegetation communities in the Mojave Desert using geostatistical techniques," *Comput. Geosci.* **26**(4), 397–410 (2000), http://dx.doi.org/10.1016/S0098-3004(99)00120-X.
17. R. G. Craig, "Autocorrelation in Landsat data," in *Proc. 13th International Symposium on Remote Sensing of Environment*, Ann Arbor, pp. 1517–1524 (1979).
18. P. M. Atkinson, E. Pardo-Igúzquiza, and M. Chica-Olmo, "Downscaling cokriging for super-resolution mapping of continua in remotely sensed images," *IEEE Trans. Geosci. Rem. Sens.* **46**(2), 573–580 (2008), http://dx.doi.org/10.1109/TGRS.2007.909952.
19. M. J. Pringle, M. Schmidt, and J. S. Muir, "Geostatistical interpolation of SLC-off Landsat ETM + images," *ISPRS J. Photogram. Rem. Sens.* **64**(6) 654–664 (2009), http://dx.doi.org/10.1016/j.isprsjprs.2009.06.001.
20. F. Chica-Olmo and M. Abarca-Hernandez, "Radiometric coregionalization of Landsat TM and SPOT HRV images," *Int. J. Rem. Sens.* **19**(5), 997–1005 (1998), http://dx.doi.org/10.1080/014311698215838.
21. M. A. Oliver, R. Webster, and K. Slocum, "Filtering SPOT imagery by kriging analysis," *Int. J. Rem. Sens.* **21**(4), 735–752 (2000), http://dx.doi.org/10.1080/014311600210542.
22. N. A. C. Cressie, *Statistics for Spatial Data*, John Wiley & Sons, New York (1993).
23. A. J. Plaza and C. I. Chang, *High Performance Computing in Remote Sensing*, Chapman & Hall/CRC Computer, Boca Raton (2008).
24. A. Cortés, "Half-duplex dynamic data driven application system for forest fire spread prediction," in *Lecture Notes in Computer Science. High Performance Computing and Applications*, Vol. 5938, pp. 1–7, Springer (2010).
25. L. S. R. Froude, "Storm tracking with remote data and distributed computing," *Comput. Geosci.* **34**(11), 1621–1630 (2008), http://dx.doi.org/10.1016/j.cageo.2007.11.004.
26. J. Le Moigne, W. J. Campbell, and R. F. Cromp, "An automated parallel image registration technique based on the correlation of wavelet features," *IEEE Trans. Geosci. Rem. Sens.* **40**(8), 1849–1864 (2002), http://dx.doi.org/10.1109/TGRS.2002.802501.
27. M. K. Dhodhi et al., "D-ISODATA: a distributed algorithm for unsupervised classification of remotely sensed data on network of workstations," *J. Parallel Distributed Comput.* **59**(2), 280–301 (1999), http://dx.doi.org/10.1006/jpdc.1999.1573.

28. S. Wang et al., "Simple grid toolkit: enabling geosciences gateways to cyberinfrastructure," *Comput. Geosci.* **35**(12), 2283–2294 (2009), http://dx.doi.org/10.1016/j.cageo.2009.05.002.
29. National Aeronautics and Space Administration (NASA), Landsat Program http://landsat.gsfc.nasa.gov.
30. A. C. Newton et al., "Remote sensing and the future of landscape ecology," *Prog. Phys. Geogr.* **33**(4), 528–546 (2009), http://dx.doi.org/10.1177/0309133309346882.
31. X. Pons et al., "Ten years of local water resource management: integrating satellite remote sensing and geographical information systems," *Eur. J. Rem. Sens.* **45**(1), 317–332 (2012), http://dx.doi.org/10.5721/EuJRS20124528.
32. L. Martínez et al., "Atmospheric correction algorithm applied to CASI multi-height hyperspectral imagery," in *Proc. RAQRS II Valencia*, pp. 170–173 (2006).
33. R. E. Wolfe, D. P. Roy, and E. Vermote, "MODIS land data storage, gridding, and compositing methodology: level 2 grid," *IEEE Trans. Geosci. Rem. Sens.* **36**(4), 1324–1338 (1998), http://dx.doi.org/10.1109/36.701082.
34. X. Pons, "MiraMon. Geographical information system and remote sensing software," Center for Ecological Research and Forestry Applications, CREAF, (2000), http://www.creaf.uab.cat/MiraMon (7 January 2013).
35. V. Palà and X. Pons, "Incorporation of relief into geometric corrections based on polynomials," *Photogram. Eng. Rem. Sens.* **61**(7), 935–944 (1995).
36. X. Pons and L. Solé-Sugrañes, "A simple radiometric correction model to improve automatic mapping of vegetation from multispectral satellite data," *Rem. Sens. Environ.* **48**(2), 191–204 (1994), http://dx.doi.org/10.1016/0034-4257(94)90141-4.
37. http://opengis.uab.cat/wms/satcat/index.htm (7 January 2013).
38. A. Zabala et al., "JPEG2000 encoding of images with NODATA regions for remote sensing applications," *J. Appl. Rem. Sens.* **4**(1), 041793 (2010), http://dx.doi.org/10.1117/1.3474978.
39. R. Alamús, J. Talaya, and I. Colomina, "The SISA/0: ICC experiences in airborne sensor integration," in *Joint Workshop of ISPRS WG I/1, I/3 and IV/4*, Hannover (1999).
40. Warehouse Inventory Search Tool, National Aeronautics and Space Administration (NASA), https://wist.echo.nasa.gov/wist-bin/api/ims.cgi?mode=MAINSRCH&JS=1 (26 May 2011).
41. http://www.kakadusoftware.com (7 January 2013).
42. JPEG2000 Part I: ISO/IEC/15444-1:2004, Part II: ISO/IEC/15444-2:2004, Part III: ISO/IEC/15444-3:2004, Geneva, Switzerland (2004).
43. F. Aulí-Llinàs et al., "J2K: introducing a novel JPEG 2000 coder," *Proc. SPIE* **5960**, 1763–1773 (2005), http://dx.doi.org/10.1117/12.633226.
44. P. Goovaerts, *Geostatistics for Natural Resources Evaluation*, Oxford University Press, New York (1997).
45. G. Matheron, *Traité de Géostatistique Appliquée*, Editions Technip, Paris (1962).
46. M. A. Oliver and R. Webster, "Kriging: a method of interpolation for geographical information systems," *Int. J. Geograph. Inf. Sci.* **4**(3), 313–332 (1990), http://dx.doi.org/10.1080/02693799008941549.
47. C. E. Woodcock, A. H. Strahler, and D. L. B. Jupp, "The use of variograms in remote sensing: I. Scene models and simulated images," *Rem. Sens. Environ.* **25**(3), 323–348 (1988), http://dx.doi.org/10.1016/0034-4257(88)90108-3.
48. C. D. Lloyd, *Local Models for Spatial Analysis*, CRC Press, Boca Raton (2007).
49. L. Pesquer, A. Cortés, and X. Pons, "Parallel ordinary kriging interpolation incorporating automatic variogram fitting," *Comput. Geosci.* **37**(4), 464–473 (2011), http://dx.doi.org/10.1016/j.cageo.2010.10.010.
50. Anonymous, "Introduction to MPI," *Int. J. Supercomput. Appl. High Perform. Comput.* **8**(3–4), 169–173 (1994), http://dx.doi.org/10.1177/109434209400800302.
51. I. Foster, *Designing and Building Parallel Programs. Concepts and Tools for Parallel Software Engineering*, Addison-Wesley (1995).

52. P. Treitz and P. Howarth, "High spatial resolution remote sensing data for forest ecosystem classification: an examination of spatial scale," *Rem. Sens. Environ.* **72**(3), 268–289 (2000), http://dx.doi.org/10.1016/S0034-4257(99)00098-X.
53. M. Casas, R. Badia, and J. Labarta, "Automatic analysis of speedup of MPI applications," in *Proc. of the 22nd ACM International Conf. on Supercomputing*, Island of Kos, Greece, pp. 349–358 (2008).

Lluís Pesquer received his BS degree in physics in 1994 at the Universitat de Barcelona (UB) and a MS degree in computational science in 2009 at the Universitat Autònoma de Barcelona (UAB). He is researcher at CREAF since 2000 and he is currently the main developer of analysis tools, remote sensing modules and geodesy libraries of MiraMon GIS software. His main research interests are: spatial analysis, geostatistics, parallel computing in geoprocessing environments, remote sensing and applied geodesy to GIS.

Xavier Pons received his BS degree in biology in 1988, a MS degree in botany in 1990, a MS degree in geography in 1995, and a PhD degree in remote sensing and GIS in 1992, all from the Universitat Autònoma de Barcelona (UAB). His main work has been done in radiometric and geometric corrections of satellite imagery, in cartography of ecological and forest parameters from airborne sensors, in studies of the spectral response of Mediterranean vegetation and in GIS development, both in terms of data structure and organization and in terms of software writing. He has recently worked in descriptive climatology models, in modeling forest fire hazards and in analysis of landscape changes from long series of satellite images. He is Full professor at the Department of Geography of the Autonomous University of Barcelona and coordinates research activities in GIS and Remote Sensing.

Ana Cortés received both her first degree and her PhD in computer science from the Universitat Autònoma de Barcelona (UAB), Spain, in 1990 and 2000, respectively. She is currently associate professor of the Computer Architecture and Operating Systems Department at UAB. Her areas of interest are high performance computing, distributed and grid computing and also scheduling and load balancing in parallel systems. Her current research interest is focused on high performance computing applied to dynamic data driven applications systems for forest fire spread prediction.

Ivette Serral received her degrees in environmental sciences, MSc in remote sensing and GIS, both at the Universitat Autònoma de Barcelona (UAB), she is researcher at CREAF since 2005. The main objective of her work is to develop and manage GIS for the public administration: the Catalan marine and coastal information system, the Catalan healthcare information system, the Andorran environmental information system, etc. She developed new methodological tools in reference systems integration and multicriteria studies and Web geodata portals. She has been collaborating with secondary schools to disseminate GIS science and applications.

Capítol 5

A geostatistical approach for selecting the highest quality MODIS daily images

Aquest capítol és una reproducció de Pesquer L., Domingo C., Pons X. (2013) “A Geostatistical Approach for Selecting the Highest Quality MODIS Daily Images” Springer *Lecture Notes in Computer Science Series*, 7887 LNCS, 608–615.

A Geostatistical Approach for Selecting the Highest Quality MODIS Daily Images

Lluís Pesquer[1], Cristina Domingo[2], and Xavier Pons[2]

[1] CREAF, Cerdanyola del Vallès 08193, Spain
l.pesquer@creaf.uab.cat
[2] Geography Department. Ed. B. Universitat Autònoma de Barcelona. 08193 Bellaterra, Spain
{cristina.domingo,xavier.pons}@uab.cat

Abstract. The aim of this work was to develop a new methodology for automatic selection of the highest quality MODIS daily images, MOD09GA Surface Reflectance product. The methodology developed here complements the quality assessment of MODIS products with a geostatistical analysis of spatial pattern images based on variogram tools. The resulting selection is formed by 26 high-quality images (from an initial dataset of 365) from throughout 2007. Most images with geometric distortion problems, such as the bow-tie effect, were rejected. The automatic selection was validated by comparing it to manual selection, which showed that it achieved an overall accuracy of 71.4%.

Keywords: MODIS, quality assessment, variogram analysis, automatic image selection.

1 Introduction

MODIS (MODerate resolution Imaging Spectroradiometer) is the most used multispectral sensor for medium spatial resolution remotely sensed images thanks to the free distribution of a wide range of products at different processing levels [1]. Each platform (on the Terra and Aqua spacecraft) views the Earth's entire surface every one to two days. This fine time resolution is used for a large number of remote sensing applications, especially when, for example, a Landsat 16-day repeat cycle is not enough for frequent monitoring purposes [2] or when cloud covered regions are often present [3]. Indeed, this high time frequency implies that a portion of daily data is unhelpful due to meteorological conditions, mainly clouds and cloud shadows. Moreover, the MODIS double-sided scan mirror and the wide swath of 2330km cause some distortions at the image edge in the along-track direction, known as a bow-tie effect, so that some scenes require quality selection [4].

In this work the quality of a complete year series of MOD09GA Surface Reflectance Daily L2G Global 500 m and 1 km product was analyzed [5]. A geostatistical approach is proposed in order to complement the different quality products available for selecting the highest quality data set. MOD09GA products provide a set of quality bands and masks. These bands are very useful when it is necessary to reject images to obtain a high quality data set. The most frequent artifacts or errors are due to radiometric

J.M. Sanches, L. Micó, and J.S. Cardoso (Eds.): IbPRIA 2013, LNCS 7887, pp. 608–615, 2013.

processing, atmospheric (aerosol) or meteorological conditions (snow, clouds, cloud shadows, etc.), and natural phenomena (*e.g.* fires). However, no quality band is provided for other problems such as some geometric distortions. This is the case of the above mentioned bow-tie effect, which is not directly considered in the quality set, and therefore an algorithm needs to be applied to remove it [6]. The present work offers a new geostatistical methodology based on variogram analysis for improving image selection.

Geostatistical methodologies have been applied to remote sensing images for different purposes [7], and they can provide parameters for describing and quantifying spatial patterns [8], measurements for spatial autocorrelation [9], procedures for downscaling images [10], tools for estimating continuous variables [11] and image classification [12]. The present work uses powerful geostatistical tools to define a representative spatial pattern to determine images with divergent spatial behavior that can be discarded. The selection methodology presented is based on this geostatistical criteria and includes a pre-selection stage using MODIS quality assessment products. This is a new and advanced approach that improves selections based only on the MODIS quality bands or masks, and results in a robust, high-quality series of images.

The MODIS data set resulting from the present work can be used in different applications, such as dynamic state vegetation [13], monitoring droughts [14], map classification [15], generating reference values for radiometric correction of other remote sensing images, such as Landsat [16], etc.

2 Materials and Methods

A complete year, 365 images, of Daily Surface Reflectance MODIS Terra product (MOD09GA), was used in this work. The study area corresponds to a selected region of the MODIS tile, h18 v04. These images were downloaded from the NASA Earth Observing System Data and Information System through the REVERB Data and Service Access Client [17]. The study region is located in Catalonia, a region of approximately 32000 km^2 in the northeast Iberian Peninsula, at the southwest extreme of Europe (Fig. 1). The extension was matched to the 197-031 path-row in the Worldwide Reference System-2 (WRS-2) Landsat tiling system so that a common study region could be used in future combined methodology studies on both sensors.

Fig. 1. Location of the study area

A first image pre-selection stage was generated by excluding sea regions from a vector mask and applying six MODIS masks included in MOD09GA: *Cloud state,*

Cloud shadow, Cirrus detected, Fire Flag, MOD35 Snow and *Internal Snow Flag*. These masks include different description states for each pixel. In this study, only one category from each mask was used. For the *Cloud state* mask, the category "clear" was selected. For the other five masks, the category "no/none" was selected (no shadow, no cirrus, no fire, no snow). Figure 2 shows the histogram of the number of valid pixels per image and the selected threshold (75% of valid land pixels, 17114 pixels at 1 km resolution). Figure 3 compares an image with 93% valid pixels (only the sea region is excluded) and an image with under 75% (65%) valid pixels.

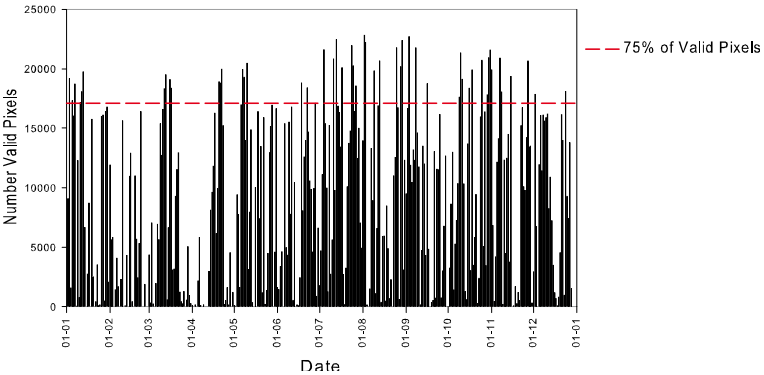


Fig. 2. Histogram of the number of daily valid land pixels

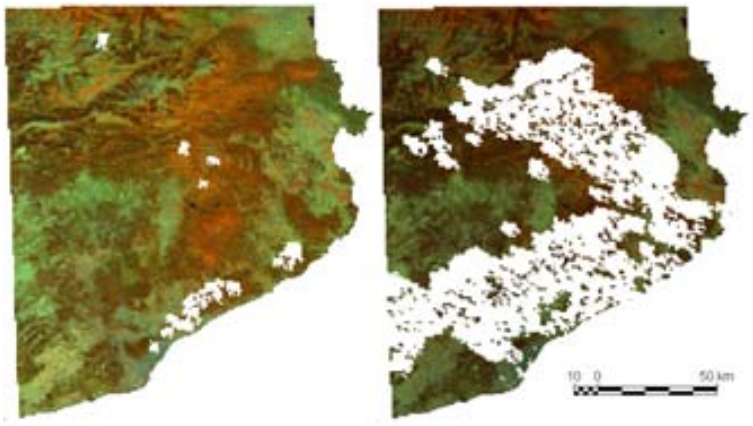


Fig. 3. Left, low (2%) number of invalid (white) pixels (scene 901). Right, 35% invalid pixels, (scene 725, image not selected).

The second step involved carrying out a geostatistical approach based on variogram analysis. The variogram, defined in Equation 1 (h is the distance between pairs of pixels and z the reflectance values), plots the dependence of the spatial variance so that the spatial pattern can be analyzed. The sampled variogram can be modeled and fitted by a continuous function to identify structural parameters that characterize the spatial pattern for any quantitative variable distribution, including, of course,

those of remotely sensed images, as in other previous studies, *e.g.* [18] (Fig. 4). These parameters are:

$$\gamma(h) = \frac{1}{2 \cdot n} \sum_{i=1}^{n} \left[z(\vec{x}_i) - z(\vec{x}_i + h) \right]^2 \quad (1)$$

• Nugget: this is the variance near the origin and represents the component of the non-spatially correlated error.

• Range: this is the distance at which the variogram reaches saturation. It represents the limit distance of autocorrelation.

• Sill: this is the variance at variogram saturation.

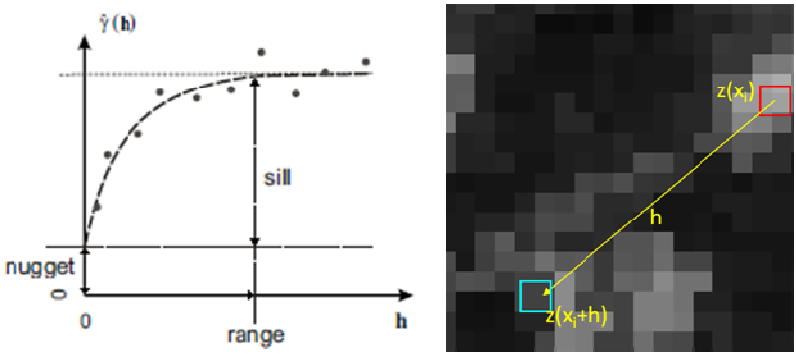


Fig. 4. On the left, variogram structure parameters. On the right, example of a pair calculation.

In this second step (Fig. 5), the images with a higher number of valid land pixels (95% in this case depending on the study region and frequently on cloud cover events) were selected and were geostatistically analyzed. Their characteristic spatial pattern was defined by a representative variogram, which is the key for rejecting images with different types of problems and distortions.

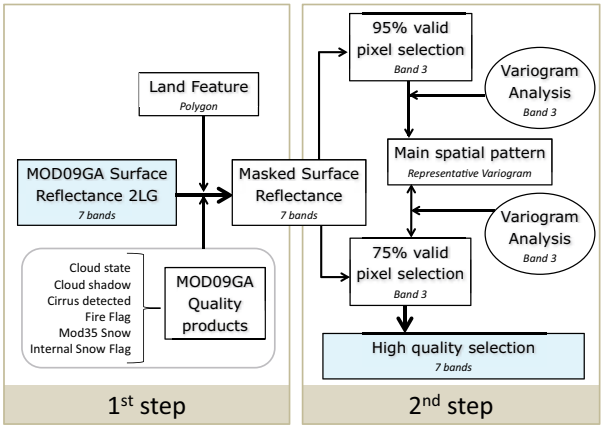


Fig. 5. Process workflow

The final step involved a massive variogram analysis of 75% valid pixel images and each of the individual variograms was compared with the representative variogram. New thresholds for the variogram structure parameters (sill and range dispersion from the representative variogram) were used in order to select or reject each image.

The *MiraMon* GIS and Remote Sensing software [19] was the main software used during this work. The authors also developed specific auxiliary tools for adapting some automatic processes to this system.

3 Results and Discussion

The selection of the images with 95% valid land pixels is shown in Table 1 (8 images) and corresponds to the variogram plots in Figure 6. This table also shows variogram structure parameters for band 3. A complete analysis for all bands is not necessary because the quality masks used are not spectrally dependent and although the radiometric range of each band is different, the spatial pattern behavior is very similar [20], except when a specific band has some radiometric problems, *e.g.* band 5. MODIS. Band 5 is affected by stripe noise, which reduces the data quality and which is therefore a problem in image analysis [21].

As explained above, these eight images were used to define a central variogram that represents a model spatial pattern.

Table 1. Variogram parameters for images with at least 95% of valid pixels

MMDD	No. valid pixels	Nugget	Range	Sill
0715	22461	0.33	20436	3.51
0726	21948	0.30	24817	5.19
0804	22818	0.07	21045	3.70
0805	22240	0.44	21130	3.26
0827	21790	0.08	17875	2.44
0901	22359	0.52	19112	2.80
0906	22717	0.45	17428	2.21
0911	21785	0.33	21901	2.32
Mean (M)	-	0.32	20468	3.18
M-2σ	-	-0.02	15715	1.55
M+2σ	-	0.65	25221	5.44

A 51 image dataset pre-selection (75% valid pixel) was compared with the representative variogram. The criterion for a definitive image selection is that the three structure variogram parameters (nugget, sill and range) are within a range that is twice the standard deviation from the mean calculated from the representative variogram. The following 26 selected images satisfy this criterion: (MMDD 2007) 105, 107, 113, 424, 624, 706, 713, 715, 726, 804, 805, 815, 827, 901, 906, 911, 1013, 1014, 1019, 1021, 1028, 1102, 1104, 1110, 1111 and 1118.

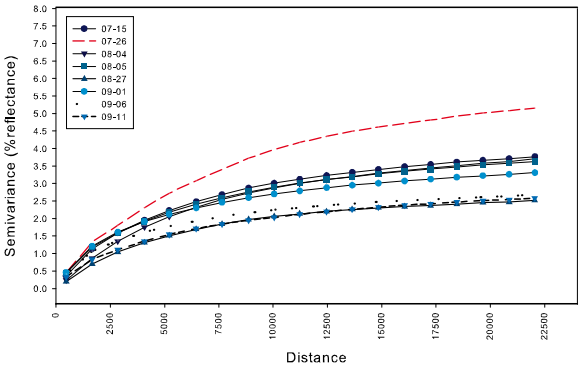


Fig. 6. Variogram plot for selected images with at least 95% of valid pixels (n=8)

This final selection was validated by comparing it with the image selection resulting from an independent manual classification. The methodology developed here showed 71.4% overall accuracy. The two methods agreed on 35 images (19 images were rejected and 16 images selected). However, in 14 cases the decision was different, four manual valid images were rejected by the present methodology (two cases were very close to the 2σ threshold), and 10 images that were rejected by the manual classification were considered valid by the present methodology.

An interesting case to illustrate the present methodology is the extra analysis applied to images with geometric distortions due to the bow-tie effect. The rejection of these distorted images, even though they have almost no cloud cover, is an excellent improvement on the selection based only on mask quality products. Figure 7 shows a zoom view of the 11 August scene (811) (with a bow-tie effect distortion that implies the loss of spatial resolution) compared to the 11 September scene (911). Note that some regions in the left scene look like their spatial resolution is nearer to 1 km instead of the nominal resolution of 500 m. Compared with the representative variogram in Figure 6, the variogram 811 has a range of 37199 m (Fig.8), this means that it is a more autocorrelated image with a less realistic spatial detail. However, the present methodology does not always detect this type of distortion, *e.g.* the scene 815 is an unusual and unresolved case that needs to be solved in future work.

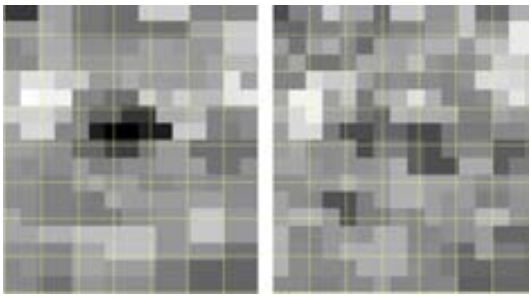


Fig. 7. Visual comparison of an image with a bow-tie effect (left) and a high quality image (center). Spatial reference yellow grid of 1 km is superposed.

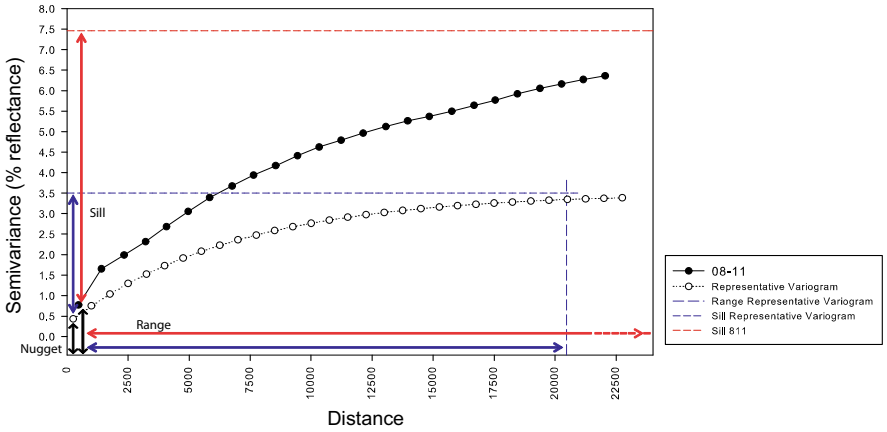


Fig. 8. Variogram of distorted image. The range is larger than the representative model

4 Conclusions

The geostatistical approach presented achieves the main objectives in order to replace the manual (visual) methodology for selecting high quality MODIS daily images and is an improvement on the automatic selection based only on MODIS quality bands or masks.

The accuracy obtained encourages the authors to develop further applications for defining pseudo-invariant areas with coherent reflectance values as a reference for reflectance model calibration of other sensors.

Acknowledgment. This work was partially supported by the Catalan Government under Grant SGR2009-1511 and by the Spanish Ministry of Science and Innovation under Grant CGL2012-33927 and by an FPU scholarship (AP2008-2016). Xavier Pons is a recipient of an ICREA Academia Excellence in Research Grant (2011-2015).

References

1. Vermote, E.F., Kotchenova, S.Y.: MOD09 (Surface Reflectance) User'sGuide, `http://modis-sr.ltdri.org/products/MOD09_UserGuide_v1_2.pdf`
2. Gao, F., Masek, J., Schwaller, M., Hall, F.: On the Blending of the Landsat and MODIS Surface Reflectance: Predicting Daily Landsat Surface Reflectance. IEEE Transactions on Geoscience and Remote Sensing 44(8) (2006)
3. Hilker, T., Wulder, M.A., Coops, N.C., Seitz, N., White, J.C., Gao, F., Masek, J.G., Stenhouse, G.: Generation of dense time series synthetic Landsat data through data blending with MODIS using a spatial and temporal adaptive reflectance fusion model. Remote Sensing of Environment 113, 1988–1999 (2009)
4. Xie, Y., Xiong, X., Qu, J.J., Che, N., Summers, M.E.: Impact analysis of MODIS band-to-band registration on its measurements and science data products. International Journal of Remote Sensing 32(16), 4431–4444 (2011)

5. Justice, C.O., Townshend, J.R.G., Vermote, E.F., Masuoka, E., Wolfe, R.E., Saleous, N., Roy, D.P., Morisette, J.T.: An overview of MODIS Land data processing and product status. Remote Sensing of Environment 83(1-2), 3–15 (2002)
6. Gómez-Landesa, E., Rango, A., Bleiweiss, M.: An algorithm to address the MODIS bowtie effect. Canadian Journal of Remote Sensing 30(4), 644–650 (2004)
7. Curran, P.J., Atkinson, P.M.: Geostatistics and remote sensing. Progress in Physical Geography 22(1), 61–78 (1998)
8. Garrigues, S., Allard, D., Baret, F.: Using First- and Second-Order Variograms for Characterizing Landscape Spatial Structures from Remote Sensing Imagery. IEEE Transactions on Geoscience and Remote Sensing 45(6), 1823–1834 (2007)
9. Craig, R.G.: Autocorrelation in Landsat data. In: Proc. of the 13th International Symposium on Remote Sensing of Environment, pp. 1517–1524 (1979)
10. Atkinson, P.M., Pardo-Igúzquiza, E., Chica-Olmo, M.: Downscaling Cokriging for Super-Resolution Mapping of Continua in Remotely Sensed Images. IEEE Transactions on Geoscience and Remote Sensing 46(2), 573–580 (2008)
11. Pringle, M.J., Schmidt, M., Muir, J.S.: Geostatistical interpolation of SLC-off Landsat ETM+ images. ISPRS Journal of Photogrammetry and Remote Sensing 64(6), 654–664 (2009)
12. Atkinson, P.M., Naser, D.K.: A geostatistically weighted k-NN classifier for remotely sensed imagery. Geographical Analysis 42(2), 204–225 (2010)
13. Zhang, X., Friedl, M.A., Schaaf, C.B., Strahler, A.H., Hodges, J.C.F., Gao, F., Reed, B.C., Huete, A.: Monitoring vegetation phenology using MODIS. Remote Sensing of Environment 84(3), 471–475 (2003)
14. Wan, Z., Wang, P., Li, X.: Using MODIS Land Surface Temperature and Normalized Difference Vegetation Index products for monitoring drought in the southern Great Plains, USA. International Journal of Remote Sensing 25(1), 61–72 (2004)
15. Bagan, H., Wang, Q., Yang, Y., Yasuoka, Y., Bao, Y.: Land cover classification using moderate resolution imaging spectrometer-enhanced vegetation index time-series data and self-organizing map neural network in Inner Mongolia. China Journal of Applied Remote Sensing 1(1) (2007)
16. Feng, M., Huang, C., Channan, S., Vermote, E.F., Masek, J.G., Townshend, J.R.: Quality assessment of Landsat surface reflectance products using MODIS data. Computers & Geosciences 38, 9–22 (2012)
17. NASA's Earth Observing System Data and Information System, http://reverb.echo.nasa.gov/reverb/
18. Woodcock, C.E., Strahler, A.H., Jupp, D.L.B.: The use of variograms in remote sensing: I. Scene models and simulated images. Remote Sensing of Environment 25, 323–348 (1988)
19. Pons, X.: MiraMon. Geographical Information System and Remote Sensing Software, CREAF (2000), http://www.creaf.uab.cat/MiraMon
20. Pesquer, L., Cortés, A., Serral, I., Pons, X.: Spatial pattern alterations from JPEG2000 lossy compression of remote sensing images: Massive variogram analysis in High Performance Computing. Journal of Applied Remote Sensing 7(1), 073595 (2013)
21. Wang, R., Zeng, C., Li, P., Shen, H.: Terra MODIS band 5 Stripe noise detection and correction using MAP-based algorithm. In: 2011 International Conference on Remote Sensing, Environment and Transportation Engineering, pp. 8612–8615 (2011)

7. Resum dels resultats

L'estructuració en subseccions d'aquest capítol de resum de resultats s'ha realitzat en funció del contingut dels anteriors i les temàtiques tractades, així com de la seva interconnexió i unitat temàtica exposades en la introducció. Els resultats s'indiquen amb una breu explicació que en recull l'essència; per conèixer-los amb profunditat, així com la metodologia a través de la qual s'han obtingut, cal consultar la publicació del capítol corresponent.

7.1 Automatització

La recerca per tal de possibilitar automatitzar els processos inherents a determinades tasques complexes en l'àmbit de l'anàlisi espacial i de la Teledetecció és un del pilars principals d'aquesta tesi. A continuació exposem els resultats dels diferents capítols per després analitzar les aportacions globals en aquest àmbit:

- Capítol 2:
    - **Ajust automàtic del variograma**: Als punts del variograma empíric s'ajusten diferents funcions no-lineals pel mètode Levenberg-Marquardt i es determina el més adequat segons un criteri estadístic prèviament escollit. Aquest automatisme és clau per poder realitzar una interpolació *kriging* d'elevada qualitat basada en el millor variograma possible sense intervenció de l'usuari. Aquesta aproximació també possibilita la seva execució en entorns informàtics distribuïts, alhora que permet una interpolació a temps real si l'entorn computacional està adequadament dimensionat.
- Capítol 3:
    - **Cerca de l'interpolador òptim:** L'execució automatitzada dels diferents mètodes d'interpolació i/o regressió amb un rang ampli de valors dels seus paràmetres i/o variables independents (pel cas de la regressió) i la posterior comparació, també automàtica, dels RMS (*root mean square*) o dels coeficients de determinació R^2 (pel cas de la regressió) allotjats

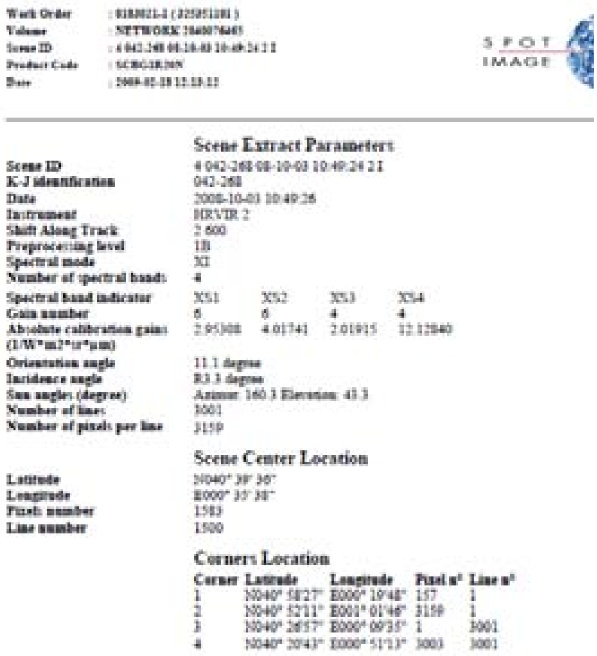
en les metadades de qualitat del resultats, permeten triar l'interpolador òptim per a cada nova mostra de dades descarregades.

- **Publicació automàtica de mapes continus:** És possible generar automàticament al servidor de mapes, amb el rigor adequat, les capes pre-preparades segons les especificacions dels estàndards de l'Open Geospatial Consortium (OGC) un cop seleccionats els millors resultats a partir del mètode descrit al resultat anterior.

- Capítol 4:
  - **Ajust automàtic del variograma**: S'ha adaptat el mètode desenvolupat en el capítol 2 per a imatges de Teledetecció de diversos tipus i resolucions espacials. La generació del variograma empíric a partir d'imatges de Teledetecció implica habitualment el tractament d'un molt més gran volum de dades i, per tant, més temps de càlcul.
- Capítol 5:
  - **Filtre automàtic d'imatges d'elevada qualitat:** Mitjançant l'ús adequat de les màscares de qualitat de les imatges de Teledetecció, la determinació del seu variograma i comparació respecte un variograma model, es trien, sense intervenció manual, les imatges de més qualitat, la qual cosa és molt important en l'elecció de les dades a tractar en grans aplicacions com, per exemple, les sèries temporals d'imatges.
- Capítol 6 (o annex 1):
  - **Generació automàtica d'àrees pseudoinvariants (PIA):** Una vegada eliminats els fenòmens ocasionals (núvols, neus, focs, distorsions geomètriques, etc) gràcies a l'aplicació d'un filtre automàtic d'imatges d'elevada qualitat que s'aplica sobre llargues sèries d'imatges diàries (10 anys en el nostre cas), és possible generar, també automàticament, zones amb valors de reflectància quasi constant, o sigui àrees pseudo-invariants.
  - **Correcció radiomètrica automàtica:** La selecció automàtica de les PIA adequades que s'usen per a una imatge concreta permet resoldre els paràmetres no fixats de les equacions del model de correcció atmosfèrica+topogràfica i, en conseqüència, realitzar la correcció radiomètrica sense la intervenció de l'usuari, la qual implicava un mètode manual d'examen visual dels valors mínims de radiàncies de la imatge i la posterior verificació de si, per la seva localització i contingut temàtic, tenien un correcte significat radiomètric.

Vistos els anteriors resultats, en primer lloc cal remarcar, la importància cabdal de les metadades en tots els processos d'automatització gràcies a un disseny força elaborat i un tractament rigorós dins el SIG (Yang *et al.* 2013). Les metadades són imprescindibles en origen, o sigui cal que els productors de dades les subministrin el més completes i estandarditzades possible. En la figura 10a, es mostra un exemple de metadades d'una imatge SPOT-4 HRVIR gens utilitzables en entorns automatitzats donat que s'han lliurat en un fitxer *pdf* preparat per a la impressió però no adequat per a una lectura identificable i estructurada. En canvi, la figura 10b mostra les metadades d'una imatge

Landsat7-ETM+ no del tot estandarditzada, però sí força més adequada com a fitxer de text pla i amb claus i valors identificables.


Work Order
Volume
Scene ID
Product Code
Date

SPOT IMAGE

Scene Extract Parameters

Scene ID	4 042-268 08-10-03 10:49:24 2 I			
K-J identification	042-268			
Date	2008-10-03 10:49:26			
Instrument	HRVIR 2			
Shift Along Track	2 600			
Preprocessing level	1B			
Spectral mode	XI			
Number of spectral bands	4			
Spectral band indicator	XS1	XS2	XS3	XS4
Gain number	6	6	4	4
Absolute calibration gains (1/W*m2*sr*µm)	2.95308	4.01741	2.01915	12.12840
Orientation angle	11.1 degree			
Incidence angle	R3.3 degree			
Sun angles (degree)	Azimut: 160.3 Elevation: 43.3			
Number of lines	3001			
Number of pixels per line	3159			

Scene Center Location

Latitude	N040° 38' 36"
Longitude	E000° 35' 38"
Pixel number	1583
Line number	1500

Corners Location

Corner	Latitude	Longitude	Pixel nº	Line nº
1	N040° 58'27"	E000° 19'48"	157	1
2	N040° 52'11"	E001° 01'46"	3159	1
3	N040° 26'57"	E000° 09'35"	1	3001
4	N040° 20'43"	E000° 51'13"	3003	3001

Figura 10a: Metadades no automatitzables en format pdf. Font: SPOT Image

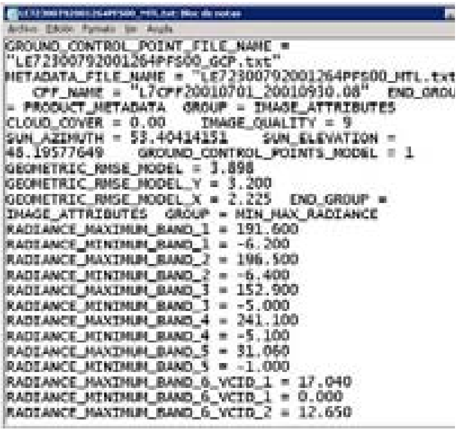

```
GROUND_CONTROL_POINT_FILE_NAME =
"LE72300792001264PFS00_GCP.txt"
METADATA_FILE_NAME = "LE72300792001264PFS00_MTL.txt
    CPF_NAME = "L7CPF20010701_20010930.08"  END_GROU
= PRODUCT_METADATA  GROUP = IMAGE_ATTRIBUTES
CLOUD_COVER = 0.00    IMAGE_QUALITY = 9
SUN_AZIMUTH = 53.40414151    SUN_ELEVATION =
48.19577649    GROUND_CONTROL_POINTS_MODEL = 1
GEOMETRIC_RMSE_MODEL = 3.898
GEOMETRIC_RMSE_MODEL_Y = 3.200
GEOMETRIC_RMSE_MODEL_X = 2.225  END_GROUP =
IMAGE_ATTRIBUTES  GROUP = MIN_MAX_RADIANCE
RADIANCE_MAXIMUM_BAND_1 = 191.600
RADIANCE_MINIMUM_BAND_1 = -6.200
RADIANCE_MAXIMUM_BAND_2 = 196.500
RADIANCE_MINIMUM_BAND_2 = -6.400
RADIANCE_MAXIMUM_BAND_3 = 152.900
RADIANCE_MINIMUM_BAND_3 = -5.000
RADIANCE_MAXIMUM_BAND_4 = 241.100
RADIANCE_MINIMUM_BAND_4 = -5.100
RADIANCE_MAXIMUM_BAND_5 = 31.060
RADIANCE_MINIMUM_BAND_5 = -1.000
RADIANCE_MAXIMUM_BAND_6_VCID_1 = 17.040
RADIANCE_MINIMUM_BAND_6_VCID_1 = 0.000
RADIANCE_MAXIMUM_BAND_6_VCID_2 = 12.650
```

Figura 10b: Metadades automatitzables en fitxer de text pla amb claus identificables. Font: USGS.

Però aquest requeriment de les metadades associades només en la font originària no és suficient, sinó que cal que els processos les usin intel·ligentment, les propaguin i les documentin de forma dirigible cap als nous resultats. Aquest ha estat el procediment funcional de les metadades en aquesta tesi, tal com demana Lynch (2008):

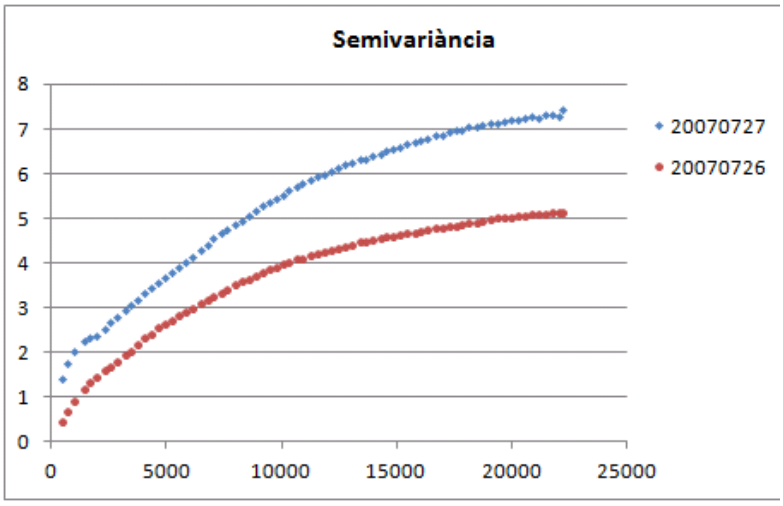
> *... defining and recording appropriate metadata — such as experimental parameters and set-up — to allow for data interpretation. This is best done when the data are captured. Indeed, descriptive metadata are often integrated within the experimental design. Description includes tracing provenance — where the data came from, how they were derived, their dependence on other data and all changes made since their capture.*

Implementar aquestes funcionalitats per a un tractament apropiat de les metadades, li correspon al SIG (Zabala and Pons 2002), i el SIG MiraMon ho ha facilitat, però també les aplicacions desenvolupades expressament en el marc de la tesi ho han explotat i respectat. La gestió rigorosa i completa de les metadades permet els resultats exposats en termes d'automatització, naturalment també aconseguits gràcies a la recerca i implementació en els algorismes adequats.

7.2 Anàlisis espacial

Els treballs en anàlisi espacial, principalment en geostadística, han produït diversos resultats que a continuació es recullen:

- Capítol 2:
  - ***Kriging* a temps real i de qualitat**: La solució paral·lelitzada en un entorn HPC i els automatismes exposats en la secció 7.1 permeten una reducció molt significativa del temps d'execució de la interpolació *kriging*, mantenint la qualitat d'un model predictiu sense discontinuïtats i generat amb totes les mostres alhora.
- Capítol 3:
  - **Automatismes d'exploració exhaustiva de l'interpolador òptim**: La recerca en automatització (vegeu secció 7.1) permet adaptar l'interpolador més adequat amb els seus paràmetres òptims a cada nou mostreig descarregat i obtenir els corresponents mapes continus amb paràmetres de qualitat.
- Capítol 4:
  - **Indicador de qualitat en la compressió**: La divergència del variograma d'una imatge comprimida amb pèrdua respecte al variograma original esdevé un indicador de qualitat que dóna més informació de l'alteració del patró espacial que l'habitual PSNR (*peak signal to noise ratio*). Aquestes divergències es mesuren a partir de les diferencies entre els paràmetres estructurals del variograma teòric de la imatge comprimida respecte de l'original (sense comprimir).
- Capítols 5 i 6 (o annex 1):
  - **Criteri addicional per la tria d'imatges:** El variograma és capaç de complementar la tria basada en màscares de qualitat en detectar anomalies geostadístiques del patró espacial de variabilitat, normalment degudes a núvols o neus parcialment no-detectades (vegeu figura 11), així com també a distorsions geomètriques inevitables en imatges de gran dallada com les MODIS.



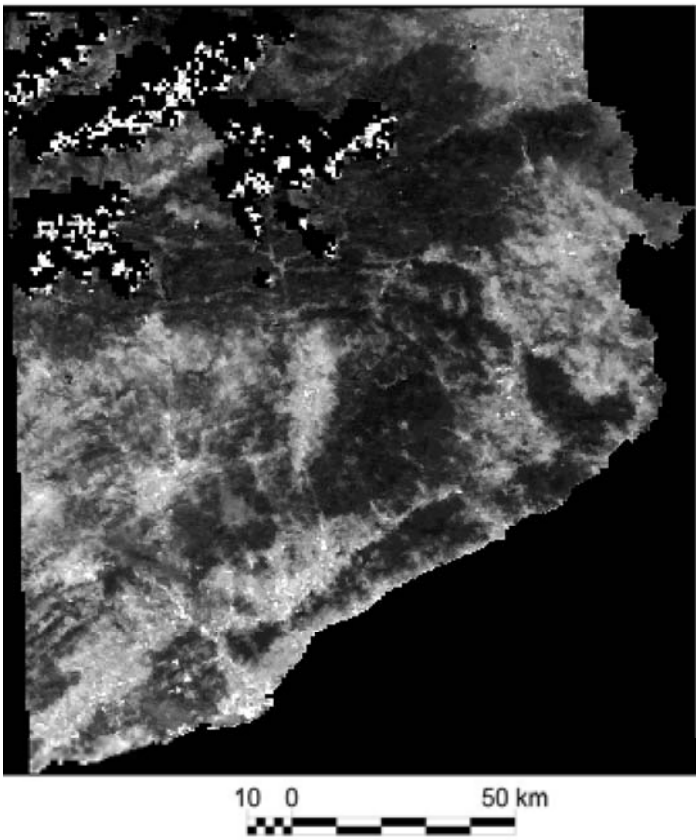


Figura 11: A la part superior es mostren dos variogrames corresponents a dues dates consecutives de diferents d'imatges Terra MODIS. En una d'elles (20070727) la màscara no ha filtrat tots els píxels amb nuvolositat i, per tant, dóna lloc a un variograma amb una variabilitat superior al variograma tipus, prèviament modelitzat. A la part inferior es mostra la banda 3 de la imatge on a l'extrem nord-oest es pot veure una zona amb força variabilitat (incorrectament emmascarada) envoltada d'una zona fosca (correctament emmascarada). L'anàlisi del variograma sí que ha detectat aquestes anomalies. Font: Elaboració pròpia.

El valor afegit d'aquesta tesi en aquest àmbit, i el tret més característic de les aportacions geostadístiques, és el seu enfocament computacional i massiu. Les anàlisis geostadístiques en Teledetecció són generades amb un nombre de píxels molt més gran que la majoria de treballs en aquest camp (vegeu Introducció) i els models interpolats són paral·lelitzats també possibilitat l'ús d'un gran nombre de mostres.

7.3 Teledetecció

Les aportacions d'aquesta tesi en l'àmbit de la Teledetecció són fonamentalment metodològiques i la principal intersecció amb la resta de temàtiques és en l'estudi del patró espacial de les imatges, sigui per avaluar mètodes de compressió o pel filtratge de les imatges de major qualitat. El darrer capítol (6) va més enllà i, encara que hereta i usa una part d'aquests anàlisis espacials, s'endinsa en els models de correccions radiomètriques, però una altra vegada amb finalitats molt coincidents amb la resta de

capítols: l'automatització i el processament avançat de grans volums de dades. El detall capítol a capítol dels principals resultats en aquest àmbit és:

- Capítol 3:
  - **Automatismes en la incorporació d'imatges de satèl·lit a models de regressió:** Tot i que essencialment ja explicat en la secció 7.1, cal afegir en la vessant de Teledetecció, que els indicadors de qualitat de les diferents possibilitats de models permeten triar en la generació de cada mapa continuo, quins productes de Teledetecció són variables independents explicatives del model de regressió i quin és el seu pes.
- Capítol 4:
  - **Variograma, indicador geospacial més complet que el PSNR:** Ja explicat el vessant més geostadístic d'aquest resultat en la secció 7.2, aquí cal remarcar que precisament l'habitual propietat d'autocorrelació espacial de les imatges de Teledetecció fa molt més interessant l'anàlisi del patró espacial amb el variograma que un indicador més global (menys especialitzat i sobre tot menys espacialitzat) com ho és el PSNR.
  - **Dimensionalitat necessària per compressió 3D d'imatges** de satèl·lit**:** Perquè els mètodes de compressió 3D millorin els mètodes 2D cal una dimensionalitat espectral prou elevada, no assolida pels sensors TM i ETM+ dels Landsat (7 bandes) però sí en certes configuracions com el cas del CASI (72 bandes) i també una sèrie temporal suficientment densa i llarga, complicada d'assolir amb Landsat (els núvols dificulten obtenir la freqüència temporal potencial) però força més viable per al Terra MODIS.
- Capítols 5:
  - **Filtre automàtic d'imatges d'elevada qualitat** Prèviament detallat aquest aspecte en la secció 7.1, en la present secció, cal concretar que les imatges de satèl·lit usades corresponen al producte diari MOD09GA de Terra-MODIS a 500 m, producte que incorpora màscares de qualitat a 500 m i 1 km de resolució espacial i en què el treball realitzat a l'article mostra millores significatives respecte d'un filtre només basat en aquestes màscares.
- Capítol 6 (o Annex 1):
  - **Correcció radiomètrica automàtica:** Explicada la seva component més automàtica, és important afegir, com a contribucions específiques en Teledetecció, l'obtenció de valors radiomètrics quasi invariants, la introducció d'un model d'atmosfera no-homogènia en funció de l'alçada, la recerca per a fer comparables radiometries entre dos series d'imatges Terra-MODIS i Landsat (MSS, TM i ETM+) amb prou significatives diferències de patrons espacials (vegeu figura 12) i, finalment, l'ús de classificacions i signatures espectrals com a calibració del model i com a criteris de qualitat.

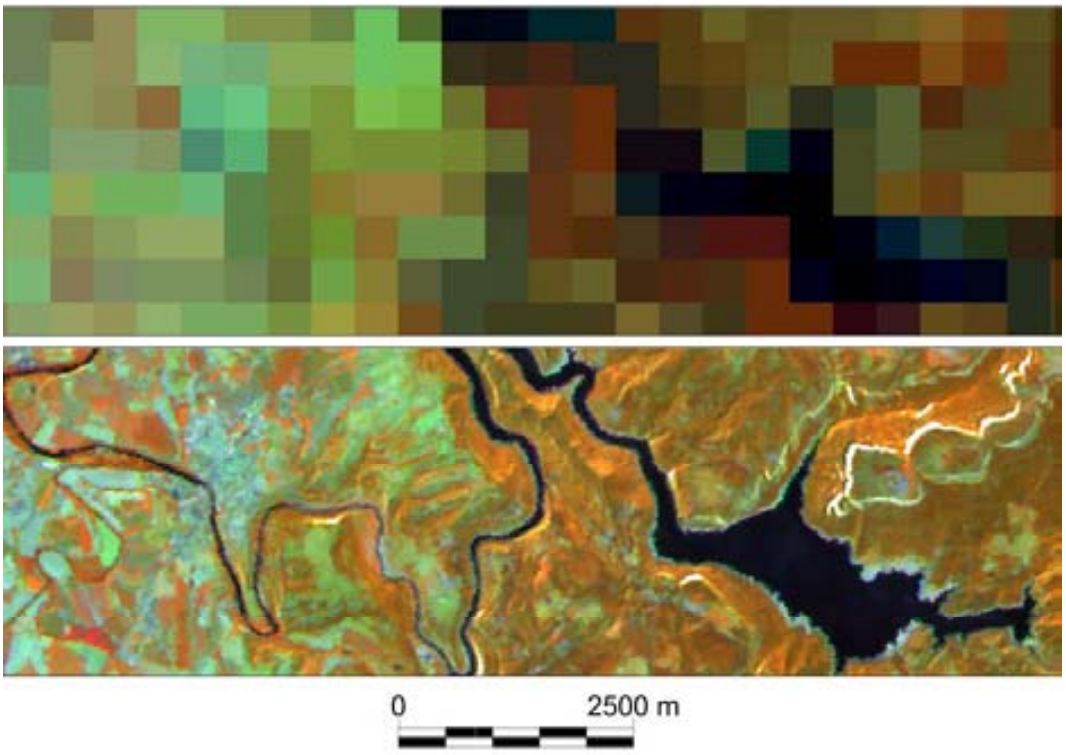


Figura 12: En la part superior, composició RGB de les bandes 2-6-1 de Terra MODIS de 500 m de costat de píxel; en la part inferior, composició RGB de les bandes 4-5-3 de Landsat-5 TM de 30 m de costat de píxel. Ambdós són productes de reflectàncies i corresponen exactament a la mateixa localització (entorn del pantà de Sau, Osona, Catalunya), en dates molt properes 29-05-2005 (MODIS) 28-05-2005 (Landsat). Font: Elaboració pròpia.

7.4 Ciència Computacional

De forma general, gran part dels resultats exposats en la secció 7.1 són també resultats interessants per a la Ciència Computacional i, per tant, no es repetiran en aquesta secció. Ara bé, cal afegir que, addicionalment, en dos capítols s'ha obert noves vies en el camp de la programació paral·lela en entorns d'elevades prestacions (HPC).

- Capítol 2:
  - **Paral·lelització del *Kriging***: El disseny de la distribució de l'esforç computacional de l'interpolador geostadístic (krigatge o *kriging*) i la seva implementació en MPI han generats resultats destacables en eficiència, avaluats en diferents entorns i conjunts de dades.
- Capítol 4:
  - **Paral·lelització de la generació del variograma**: Seguint la via de recerca iniciada en el capítol 2, s'ha dissenyat i implementat una solució de paral·lelització de generació del variograma empíric, vàlida per un ampli ventall d'imatges de Teledetecció i amb resultats rellevants des del punt de vista computacional (exemple a la figura 13).

Les aportacions d'aquesta tesi en l'àmbit computacional queden ben identificades en el llistat de citacions per part de tercers (vegeu l'annex 3) que el capítol 2 (el treball més computacional dels que componen aquesta tesi) ha aconseguit en un curt període posterior a la seva publicació.

També cal remarcar que l'estalvi molt important en temps d'execució de les aplicacions paral·lelitzades i executades en un entorn HPC ha permès ampliar el nombre de tests, realitzar variacions amb diferents paràmetres, introduint o traient variables, etc, i d'aquesta forma refinar millor alguns dels models automàtics dels altres capítols.

N Workers	Time (s)	Speedup
0	5485.03	
2	3847.47	1.43
4	1921.62	2.85
8	965.55	5.68
12	644.26	8.51
16	483.40	11.35
20	393.67	13.93
24	347.15	15.80
28	306.33	17.91
32	304.39	18.02

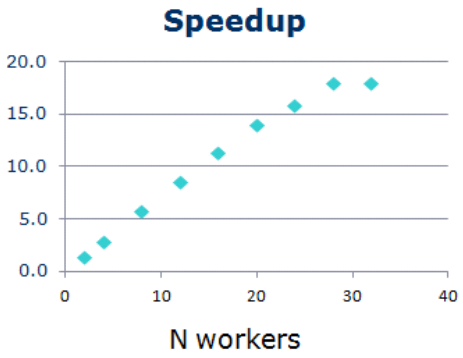


Figura 13: Resultat computacionals de la generació del variograma per a una imatge CASI de 6.5 km^2 de superfície i 3 m x 3m de costat de píxel (capítol 4). Font: Elaboració pròpia.

7.5 Compressió d'imatges

El capítol 4 és la publicació que aporta resultats en l'àmbit de la compressió d'imatges amb pèrdua. Aquests resultats han estat aplicats sobre imatges de Teledetecció i, per tant, els resultats relacionats ja han estat explicats en la secció 7.3 o, en el seu vessant més espacial, en la secció 7.2. Finalment, també aquest capítol 4 ha necessitat de les eines i la recerca en Ciència Computacional donats els grans volums de dades implicats en les seves anàlisis. Serveixi doncs aquest darrer paràgraf, com a exemple de la interrelació de les diferents temàtiques tractades durant la tesi.

8. Conclusions finals

8.1 Conclusions

La tesi que es presenta en els anteriors capítols d'aquesta memòria és un treball transversal en Teledetecció, SIG, Geostadística i Ciencia Computacional que se sosté sobre dos pilars principals: l'**anàlisi espacial** i la recerca en l'**automatització** per al tractament de grans volums de dades. La conclusió més global neix d'aquesta propietat transversal: la gestió rigorosa i eficient de grans volums de dades en el context de la Ciència de la Informació Geogràfica cal abordar-la des d'una visió àmplia, on cada disciplina hi pugui incorporar les seves potencialitats però també amb un plantejament obert que permeti aprofundir en aquells aspectes on la recerca pugui dur més enllà el nostre coneixement.

Probablement un dels resultats més rellevants de la col·laboració sinèrgica entre aquestes diferents disciplines consisteix en les aportacions en recerca sobre automatitzacions per a algorismes geostadístics. Aquesta recerca ha estat reclamada com a necessària per autors com Hengel (2009):

> *Several years ago, geostatistical analysis was considered to be impossible without intervention of a spatial analyst, who would manually fit variograms, decide on the support size and elaborate on selection of the interpolation technique.*
> *Automated mapping is still utopia for many mapping agencies.*
> *All this proves that automated mapping is an emerging research field and will receive significant attention in geography and Earth sciences in general.*

i aquesta tesi ha volgut avançar en aquesta línia.

L'automatització permet incorporar les metodologies geostadístiques a entorns computacionals d'elevades prestacions i, per tant, usar-les en aplicacions a temps real i per a l'anàlisi de grans volums de dades. La reducció, del habituals llargs temps d'execució dels mètodes geostadístics, obtinguda en un entorn d'elevades prestacions ha estat aconseguit mitjançant la distribució de la càrrega computacional amb mètodes

de paral·lelització del codi, gràcies a un disseny *master*/*worker* adequat a la gestió de les dimensions de les dades i l'enviament de missatges amb MPI (en llenguatge C) entre les unitats de processament. La solució de paral·lelització en aquests termes és innovadora i ha obtingut uns resultats en eficiència avaluats en diferents condicions.

La recerca en automatitzacions per a metodologies geostadístiques s'ha aplicat tant a dades vectorials, com a dades ràster, en aquest cas sobre imatges de Teledetecció de diferents sensors tant aeroportats com espacioportats, i amb diferents característiques espacials. L'aplicació de tècniques geostadístiques sobre grans volums de dades ràster (per exemple llargues sèries d'imatges de satèl·lit) és un camp poc explorat i, en conseqüència, els resultats aconseguits constitueixen una aportació novedosa d'aquesta tesi.

L'anàlisi geostadística d'imatges de Teledetecció de llargues sèries temporals ha permès afegir elements d'estudi del patró espacial que milloren altres tècniques que no tenen en compte la component regionalitzada de les magnituds radiomètriques de les imatges. Alguns exemples en són la selecció d'imatges d'elevada qualitat i l'estudi de les alteracions provocades pels mètodes de compressió amb pèrdua.

La selecció d'imatges Terra MODIS d'elevada qualitat ha tingut un paper determinant en la generació de polígon pseudo-invariants i els corresponents valors radiomètrics de referència. Aquests valors han permès resoldre l'aplicació del model de correcció radiomètrica de Pons and Solé-Sugrañes (1994) per a imatges Landsat (MSS, TM i ETM+) en aquestes àrees i estimar una solució per a tota una escena. D'aquesta manera, aquesta correcció (atmosfèrica+topogràfica) pot realitzar-se automàticament i, per tant, aplicar-la sobre llargues sèries temporals d'imatges. Aquests processaments de llargues sèries han de permetre un estudi acurat i exhaustiu del territori, tant des del punt de vista espacial (zones d'estudi molt extenses a una escala de treball d'elevada resolució), com de les seves dinàmiques temporals.

L'aprofitament adequat de les metadades com a clau per l'automatització dels processos és una aportació significativa d'aquesta tesi ja que dóna un paper addicional a les metadades que va més enllà de les possibles utilitats en catàlegs i les seves cerques associades, o en la construcció i gestió d'infraestructures de dades espacials (IDE). Aquest nou rol (que no és absolutament original però probablement pocs treballs li han donat un pes tant rellevant) afegeix uns trets més geogràfics que els usos habituals només aplicats a les cerques sobre catàlegs, sovint allunyades de la seva component geogràfica i tot el que pot aportar.

El SIG ha estat fonamentalment l'element integrador i de gestió geogràfica rigorosa de totes aquestes metodologies al voltant de l'anàlisi espacial i la Teledetecció. Alhora, ha estat qui ha mantingut la coherència global i el tractament acurat de les metadades. Addicionalment, en un dels treballs presentats (capítol 3) també ha estat objectiu de recerca.

Encara que la compressió d'imatges no ha estat un tema central en la tesi, la recerca que s'ha realitzat sobre l'alteració del patró espacial dels diferents mètodes estudiats conclou amb unes aportacions no menors: es proposa el variograma com a nou indicador de qualitat de la compressió amb pèrdua sobre imatges de Teledetecció, de manera que es permet identificar compressions més conservatives respecte de les propietats espacials i els patrons de variabilitat originals.

Finalment caldria preguntar-se per l'aplicabilitat de les metodologies investigades en aquesta tesi: Són específiques dels àmbits geogràfics d'estudi? Són exclusives als tipus d'imatges de Teledetecció estudiats? Respecte a l'àmbit geogràfic, no hi ha actualment cap limitació en la possibilitat d'aplicar les metodologies en altres zones d'estudi per a la majoria de treballs realitzats. La disponibilitat d'imatges de Landsat i MODIS és generalitzable a qualsevol part del món, encara que naturalment cada regió té les seves característiques (per exemple algunes zones tindran més nuvolositat que d'altres i alguns criteris de selecció d'imatges potser s'hauran d'adaptar, però no sembla que això qüestioni l'aproximació). En algun cas concret, puntual en el context del material usat en aquesta tesi, cal reconèixer, però, que serà més complicat, però l'especificitat geogràfica és una realitat. Així, per exemple, és difícil disposar d'imatges hiperspectrals com les CASI, facilitades en aquest cas per l'Institut Cartogràfic de Catalunya, o d'un seguiment de dades tan i tan complet com el relatiu a presència/absència o nombre d'individus de flora i fauna subministrat per l'Estación Biológica de Doñana. Respecte al tipus d'imatges, una gran part de la recerca (per exemple l'anàlisi de patrons espacials, la selecció d'imatges d'elevada qualitat, etc) és perfectament traslladable a imatges d'altres sensors i satèl·lits, però en algun cas caldrà realitzar adaptacions o potser desenvolupar nova recerca, ni que sigui per validar la seva universalitat, com pot ser el cas de la correcció radiomètrica automàtica per imatges de diferents tipus als aquí analitzats. Aquestes adaptacions, com per exemple les relatives a la nova configuració espectral de Landsat-8 OLI-TIRS respecte als seus predecessors, com Landsat-7 ETM+ (figura 13), obren pas a la següent secció, quines poden ser les vies futures de recerca?

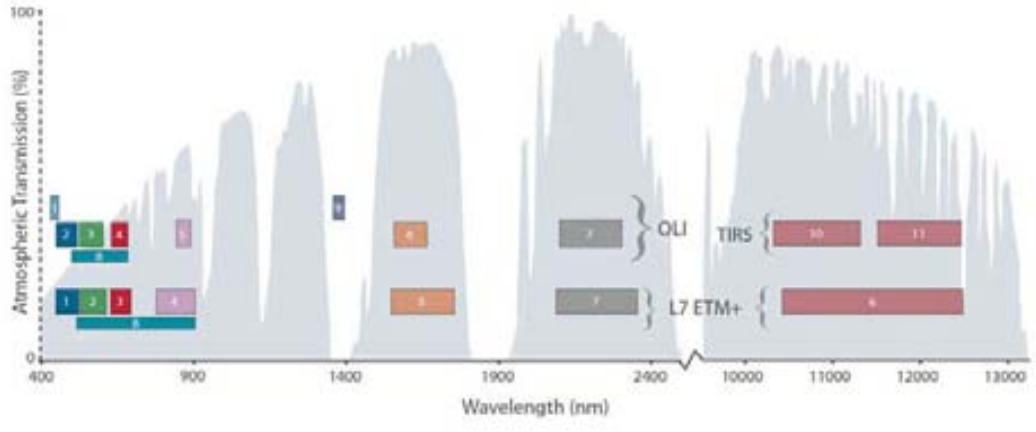


Figura 14: Gràfica comparativa de la configuració espectral de Landsat 7 ET+ i Landsat 8 OLI-TIRS. Font: Landsat Science, NASA.

8.2 Vies futures

Naturalment, la tesi no és el final d'una recerca, i com qualsevol procés intel·lectual sovint obre més interrogants que no pas els que soluciona, aspecte que cal viure com un estímul en el context de la motivació per aprendre. Algunes de les portes que han quedat obertes semblen prou interessants per continuar la recerca iniciada en aquesta tesi. Entre les propostes per a aquestes vies futures ens agradaria recollir les següents:

- Aprofundiment de les metodologies geostadístiques en l'àmbit de la Teledetecció:
    - Adaptació de les metodologies exposades a altres tipus d'imatges.
    - Detecció de patrons anòmals.
    - Caracterització de patrons de variabilitat a diferents resolucions espacials.
    - Síntesi d'imatges de diferents resolucions espacials i temporals i diferents configuracions espectrals.
    - Reompliment de zones de núvols, de zones amb errors del sensor, etc.
- Recerca en metodologies d'anàlisi espacial (incloent les geostadístiques, però no exclusivament) en models espacials predictius i en seus models d'incertesa associats, principalment per a variables climàtiques.
- Aplicabilitat i recerca de les metodologies computacionals a la gestió de gran volums de dades en el camp de la Ciències de la Informació Geogràfica (*Big Data)*.
- Recerca en l'estudi del disseny òptim per l'automatització de processos d'anàlisi espacial i de tractament (geomètric i radiomètric) d'imatges de teledetecció tenint en compte criteris com ara:
    - Qualitat del resultat.
    - Eficiència computacional.
    - Estandardització de dades i processos involucrats.

El futur no està escrit i diversos factors influiran en quins seran els camins a seguir, però entre ells el principal serà l'èxit de les propostes de projectes de recerca sotmeses a diferents convocatòries que incloguin tasques directament relacionades amb aquestes possibilitats, les quals determinaran quines línies queden en espera i quines poden abordar-se. En qualsevol cas, tant des del punt de vista temàtic com metodològic, totes elles són interessants pel grup de recerca GRUMETS, i alhora donarien continuïtat a la recerca just encetada en aquesta tesi.

Capítol 6 o Annex 1

Automatic and improved radiometric correction of Landsat imagery using reference values from MODIS surface reflectance images

Aquest capítol és una reproducció de la versió sotmesa (19-03-2014) per a primera revisió de Pons X., Pesquer L., Cristóbal J., González-Guerrero O. “Automatic and improved radiometric correction of Landsat imagery using reference values from MODIS surface reflectance images” a la revista *International Journal of Applied Earth Observation and Geoinformation.*

Automatic and improved radiometric correction of Landsat imagery using reference values from MODIS surface reflectance images

X. Pons [a], L. Pesquer [b], J. Cristóbal [c], O. González-Guerrero [a]

[a] Grumets Research Group. Dep Geografia. Edifici B. Universitat Autònoma de Barcelona. 08193 Bellaterra, Catalonia, Spain.

[b] Grumets Research Group. CREAF. Edifici C. Universitat Autònoma de Barcelona. 08193 Bellaterra, Catalonia, Spain.

[c] Geophysical Institute and Institute of Northern Engineering, University of Alaska. 903 Koyukuk Dr, Fairbanks, USA.

Abstract

This work contributes to the automatic generation of surface reflectance products in the solar bands of several types of Landsat images. These reflectance products are generated by a new approach developed from a previous simplified radiometric (atmospheric + topographic) correction model. We have kept the main characteristics of the previous model (consideration of incidence angles and cast-shadows through a Digital Elevation Model [DEM], Earth-Sun distance, etc.) and added the following characteristics to the new approach: 1/ A fitting model based on reference values from pseudoinvariant areas automatically extracted from existing reflectance products (Terra MODIS MOD09GA). This guarantees the coherence of the internal and external series, and makes unnecessary to provide extra atmospheric data for the acquisition date and time, dark objects or dense vegetation. 2/ A spatial model for the atmospheric optical depth that uses detailed DEM. 3/ A design that makes it possible to automatically process large time-series of images to produce a consistent surface reflectance product at the Landsat spatial resolution. The MODIS products are selected automatically by applying quality criteria that include a geostatistical pattern model. 4/ The approach can be applied to most images, acquired now or in the past, regardless of the processing system, with the exception of those with extremely high cloud coverage. Using Landsat as a long time-series requires to admit images having cloudy areas to properly do, for example, drought monitoring at detailed resolution. The new methodology has been successfully applied to images obtained with MSS, TM and ETM+ sensors, and also to different formats and processing systems: LPGS and NLAPS from the USGS; CEOS from ESA; and even to images with significant (60 %) cloud coverage and SLC-off. Finally, the reflectance products have been validated with some example applications: spectral signatures generation and classification, achieving results that are reasonable and very close to other reflectance products.

Keywords: Radiometric correction; Landsat; MODIS; Pseudoinvariant area.

Highlights

- Landsat atmospheric+topographic correction based on MODIS pseudoinvariant areas, PIA.
- No atmospheric data required; no dark objects or dense vegetation needed.
- Internal (Landsat) and external (MODIS) surface reflectance series coherence.
- Supports MSS, TM, ETM+, cloudy images, SLC-off and LPGS+NLAPS+ESA processing chains.
- All processes (PIA extraction, fitting of Landsat parameters, etc) are automatic.

Introduction

Landsat imagery has been collected since 1972, resulting in one of the longest continuously acquired collections of Earth observation data. It is the remotely sensed image programme that has the higher impact for scientific, management and policy-making purposes at detailed spatial resolution (Goward and Masek, 2001). As an example, Newton et al. (2009) showed that the imagery provided by the Landsat series, with its MSS, TM and ETM+ sensors, was by far the most commonly used remotely-sensed satellite imagery in Landscape Ecology studies, accounting for 42 %. The second most used was NOAA-AVHRR, which accounted for 4 %, and the third was SPOT-HRV/HRG, accounting for 3 %. In 2008, the United States Geological Survey (USGS) made access to the Landsat archive free, so that the scientific community began to use it even more (Vermote et al., 2008).

The Landsat image distribution and processing level procedures have changed since 1972 and therefore the post-processing steps, mainly geometric and radiometric correction, have also changed. In the case of the USGS, Landsat Level 1G (system correction of sensor and platform derived distortions) was the standard product until 2008, when USGS started to provide a georeferenced L1T (terrain corrected) product. Landsat Surface Reflectance products have been freely available since 2013 through Climate Data Record (CDR) (U.S. Geological Survey, 2013); it is generated from the software called Landsat Ecosystem Disturbance Adaptive Processing System (LEDAPS), based on the work of Masek et al. (2006). However, CDR product distribution from the USGS archive is currently not complete because the scenes that come from the National Landsat Archive Processing System (NLAPS) are not fully processed yet, although the list of existing reflectance products is being updated continually (see http://landsat.usgs.gov/documents/L4-5TM_NLAPS.xlsx). Moreover,

the distribution rights for scenes belonging to some regions and specific periods are assigned to other agencies, for example ESA (European Spatial Agency 2005), and this constitutes a significant lag for obtaining continuous time-series for these large regions. It is worth noting that USGS considers that the applied radiometric correction is not yet consolidated, as indicated in the Product Guide “Landsat climate data record (CDR). Surface reflectance” (U.S. Geological Survey, 2013), and therefore it is a challenge to continue carrying out research on Landsat radiometric correction.

Atmospheric conditions (water vapour, aerosols, etc.) and different illumination caused by the solar position according to the acquisition date and time, location on Earth and relief (cast shadows, etc.), can cause undesired artefacts on remote sensing images, and thus it is very important to apply the process known as radiometric correction (Pons and Solé, 1994; Richards and Jia, 2005; Janzen *et al*., 2006). Radiometric correction is a set of techniques designed to convert the digital values captured by a sensor to physical quantities of interest, such as radiance, reflectance or surface temperature (Pons and Arcalís, 2012). This transformation is required, for example, to facilitate the comparison between the same or different remote sensors at different times, as well as to compare satellite or aerial data with field-based sensors (Franklin and Giles, 1995). Radiometric correction is a prerequisite for creating high-quality scientific data (Chander et al., 2009). Among other applications, it makes it possible to discriminate between product artefacts and real changes in the Earth processes (Roy et al., 2002) and to accurately produce land cover maps and detect changes (Song et al., 2001).

In the radiometric correction process, topographic effects (incidence angles, cast shadows) can be successfully considered using an accurate source of elevation such as high resolution digital elevation models (DEM) from local cartographic institutes, or from worldwide freely-available data, like the NASA Shuttle Radar Topographic Mission (Rabus et al., 2003) or the ASTER Global Digital Elevation Map (Slater et al., 2011). Some years ago the availability of DEM in some parts of the world was a problem, but nowadays it is not, so topographic correction should be applied to any optical remote sensing product (Riaño et al., 2003; Hale and Rock, 2003; Hantson and Chuvieco, 2011). On the other hand, to account for atmospheric effects, detailed information on atmospheric parameters such as water vapour or ozone is often required. Atmospheric radiosondes at satellite pass can provide these data, but it should be taken into account that a single atmospheric radiosonde is not usually representative of the atmospheric conditions of images that cover relatively large zones, which is the case of the images provided by the Landsat, NOAA and MODIS sensors, especially in areas with highly variable relief (Cristóbal et al., 2009). The AERONET network is another

important source of atmospheric data, but as in the case of radiosonde data, its ground station distribution is not wide enough to provide atmospheric parameters over large areas (Themistocleous et al., 2012). A common drawback of these two approaches is they are not useful for older imagery. Therefore, although we acknowledge that these approaches are often considered ideal, they cannot be applied as a general method over the whole Landsat time-series.

Several proposals have been made to overcome, at least partially, these difficulties for applying radiometric correction in a general way, from the simplest dark-object methods (Chavez 1988) to other, more complete methods, integrating several factors but keeping general feasibility as a basic principle (*e.g.,* Pons and Solé, 1994, among many others).

Nevertheless, times have changed. Together with the availability of near-global, detailed DEM, the current free availability of a large part of the Landsat data bank makes it possible to contemplate a new paradigm in which medium-high resolution Earth observation data can be used as long time-series in local studies, but also at a global scale. Landsat data can contribute to understand global phenomena using a close-up approach, synergic with other data with a higher temporal revisit time, but with a spatial resolution that hardly explains some aspects in areas of landscapes being complex due to relief or to human history (Pons et al. 2014). This exciting new situation allows new scientific goals to be formulated in a variety of fields, from global change to land planning. However, when we face this possibility, we realize that although there are currently many different radiometric correction methods (Vicente-Serrano et al., 2008), this does not always guarantee radiometric homogeneity in large time-series (Schroeder et al., 2006), which shows that radiometric correction is not a completely resolved methodology (Feng et al., 2012), a fact especially true for Landsat (Masek et al., 2006).

Current methods have to be revisited in order to be automated in this new era of big data. In fact, the recent USGS initiative cited above, which aims to start providing a global product of surface reflectance, tackles the same concern that has motivated us to carry out this research. Nevertheless, we believe there is another important objective to address: to produce series that are maximally robust “internally” while being highly consistent when compared to other principal time-series of remote sensing images, such as MODIS (Potapov et al. 2008).

The aim of this work is to revise a previous radiometric (atmospheric and topographic) correction method (Pons and Solé, 1994), which has been widely used with Landsat imagery, by using rigorous metadata management that allows an automatic radiometric correction of most Landsat imagery and is highly consistent with the Terra-MODIS daily

reflectance products. The atmospheric part of the method will be improved by accounting for more realistic atmospheric conditions using pseudoinvariant areas (PIA) (Hadjimitsis et al., 2012) generated through 10-year series of Terra-MODIS imagery and polynomial fitting of atmospheric optical depth using MODTRAN (2012) simulation. Other works, as the one of Gao et al. (2010), have shown that MODIS-like data could be used as a consistent reference for Landsat-like satellite data. The method will then be applied automatically in a heterogeneous area, in terms of landscape and relief, using a set of Landsat imagery from different platforms, sensors, processing types and different levels of cloud clover. In other relevant contributions, as the one of Feng et al. (20013), only imagery practically without clouds is selected (Global Land Survey, GLS), but using Landsat as a long time-series requires to admit images having cloudy areas to properly do, for example, drought monitoring. Finally, an evaluation of surface reflectance through a spectral signature analysis and image classification will be carried out.

Before proceeding, we would like to mention that, as in the case of the CDR product (U.S. Geological Survey, 2013), we have not included Landsat-8 imagery in our approach because, at the date of submission of this paper, a year after the launch of the LDCM, USGS was in the process of reviewing the quality of the data and we believe that there will be opportunities in the near future to experiment with the new images, as well as with others (*e.g.*, Sentinel-2). However, we believe that the methodology described here could be easily applied to these new missions, as in the case of the previous methodology.

1. Model presentation

The proposed model (flowchart in Fig.1) converts from digital number (DN) values to ground reflectances on the solar spectrum at ground level based on the simplified radiometric correction model developed by Pons and Solé (1994) and designed for Landsat imagery, although it has been applied to other platforms and sensors, such as Terra/Aqua-MODIS, SPOT-HRG/HRV, IRS-LISS-III, PROBA-CHRIS or the airborne CASI (Zha et al., 2005; Serra et al., 2012; Román et al., 2005; García Millán et al., 2013; Zabala et al., 2010). This model has been used to radiometrically correct imagery for several applications, such as forest mapping (Vázquez 2008; Pérez-Cabello et al., 2010; Salvador et al., 1997; Zabala and Pons 2011), crop mapping (Barbosa et al., 1996; Serra and Pons 2008; Nuarsa et al., 2010; Moré et al., 2011), energy flux modelling (Cristóbal et al., 2011), grassland studies (Zha et al., 2003; Liu et al., 2005), climate-related modelling (Collado et al., 2002; Cristóbal et al., 2009),

biomass estimation (Lopes et al., 2009, Barrachina et al., 2010), water monitoring (Bustamante et al., 2009; Sánchez et al., 2010; Pons et al., 2012) and forest fire research (Chuvieco et al., 2002a; Chuvieco et al., 2002b; Oliveras et al., 2009), among others.

The model includes topography and atmospheric effects and takes into account several factors, such as relief and solar position (incidence angles and cast shadows), Earth-Sun distance, optical depth, exoatmospheric solar irradiance and sensor calibration parameters. The model needs two main inputs: the radiance received by the sensor from an area where there is only (or mainly) atmospheric contribution and the atmospheric optical depth. Other parameters used in the model, such as date and time of the image, are read directly from the image metadata or are easily available from external sources.

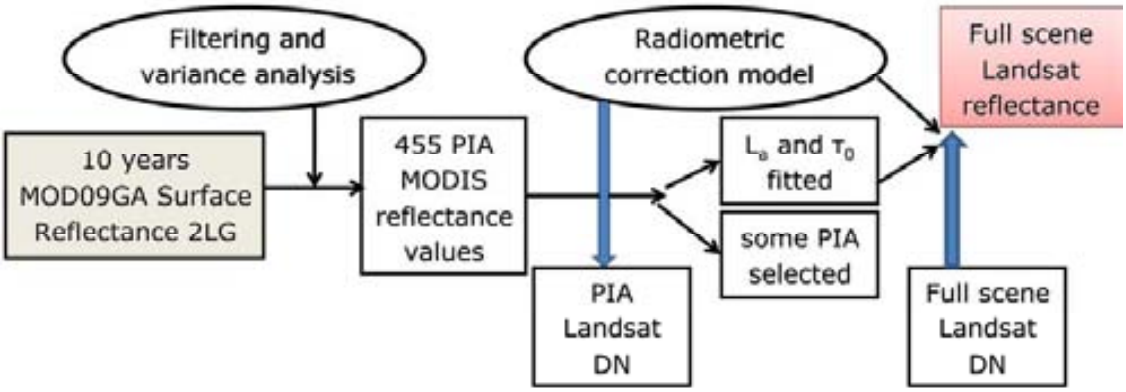


Fig. 1: Flowchart for the proposed methodology, from DN to reflectance values. DN: digital numbers; PIA: pseudoinvariant areas; 2LG

The model retrieves reflectance at ground level as the simplified radiometric correction model (Pons and Solé, 1994) in equation (1):

$$\rho = \frac{\pi \cdot [L - L_a] \cdot d^2}{\cos\theta \cdot E_0 \cdot \tau_1 \cdot \tau_2} \quad (1)$$

where ρ is the spectral reflectance at ground level, E_0 is the spectral exoatmospheric solar irradiance ($W \cdot m^{-2}$), L is the spectral radiance at sensor level, θ is the incidence angle between the solar vector and the normal vector of the terrain (accounting for its slope and aspect), d is the Sun-Earth distance in astronomical units, L_a is the spectral radiance received by the sensor from an area where there is only atmospheric contribution, τ1 is the atmospheric transmittance through the path Sun to Earth and τ2 is the atmospheric transmittance through the path Earth to Sun; τ1 and τ2 are both wavelength dependent and calculated according to equations (2) for the Sun to Earth direction and Earth to Sun, (3), respectively:

$$\tau_1 = e^{\left(\frac{\tau_0}{cos\,(s)}\right)} \quad (2)$$

$$\tau_2 = e^{\left(\frac{\tau_0}{cos\,(v)}\right)} \quad (3)$$

where τ0 is the atmospheric optical depth, s is the illumination zenith angle of the Sun and v is the view zenith angle of the sensor.

As this paper aims to improve and automate the previous radiometric correction model, we would like to explain a modification that has been applied over recent years and that is related to the "Implementation" section of the original paper. It was discussed then how to adapt reflectance values to the 8-bit per pixel and spectral band data type, which was usually used at that time; later on, the solution was to change to using real (float, 4 byte) representations (often in percentage) directly or, for storage saving purposes, using short integer (2 byte) values conveniently scaled after multiplying reflectance values by 10000. At the same time, this is a general converging solution for the new sensors with higher radiometric resolutions (for example 12 bits in the Landsat-8 Operational Land Imager, OLI, and in the Sentinel-2 Multispectral Imager, MSI). This modification was implemented in the MiraMon GIS & RS software (Pons 2000) along with other improvements not discussed here (specification of the incidence angle considered not reliable because it exceeds Lambertian limits for most surfaces, detection of cast-shadowed pixels, possibility to model the angle of incidence from a DEM of higher resolution than the image, etc).

The atmospheric part of the model is based on calculating L_a (Eq. 1) and τ_0 (Eqs 2 and 3) for a specific image date and time and for each spectral band. According to Chavez (1988) and Bariou et al. (1986), and in the original model formulation, L_a can be estimated from surfaces that do not receive any direct solar irradiation by using a procedure often referred to as dark object subtraction or the histogram minimum method. We obtained it manually by means of a histogram inspection in order to avoid problems such as false minima in the histogram (for example, due to sensor errors) or approaches based on cumulating a certain percentage of minimum values of the histogram before deciding the L_a to be used. We tried to determine a reasonable minimum in accordance with completely shadowed surfaces (self shadows and cast shadows), or with water bodies in the NIR and, especially, the SWIR. On the other hand, in the original model formulation, τ0 was considered constant for the full scene (according to, for instance, Dozier, 1989).

However, these procedures can be problematic in a variety of situations. For example, approximating L_a by simple histogram+shadows methods is not convenient in the

visible part of the spectrum when there are no hard shadows on the image. Regarding τ0, it is obvious that a very complex atmospheric situation, with many clouds, etc, over a large area is extremely difficult to model, but it is also true that in these situations optical remote sensing imagery is not used; nevertheless, assuming a completely homogeneous atmosphere can be too simple for many still-useful optical images.

For these reasons the new procedure improves the estimation of L_a and τ_0 and carries this out automatically, but keeping the original principle of being a proposal that can be applied to most images in the solar spectrum domain, acquired now or in the past, with the exception of those with very high cloud coverage. The basic idea is to use the well-known and validated Terra-MODIS time series of reflectance surface (Kotchenova and Vermote, 2007; Vermote and Kotchenova, 2008) to detect pseudoinvariant areas (PIA) and to use the reflectance on these areas to estimate L_a and τ_0 for the Landsat scenes. In other words, once a data bank of PIA has been obtained for the region of interest, and given a Landsat image in this region, L_a and τ_0 at the time of the satellite pass are estimated for each PIA by fitting equation (1) (see 3.3 . “Pseudoinvariant area (PIA) generation through MODIS imagery” for details). The atmospheric optical depth is initially modelled by means of a third order polynomial function, equation (4) for TM and ETM+ sensors and equation (5) for TM and ETM+ sensors, fitted with MODTRAN (2012) mean atmospheric optical depth values computed in transmittance mode with several standard atmospheres (US Standard 1976, MidLattiude Summer, MidLattitude Winter, SubArtic Summer, SubArticWinter and Tropical) and discrete altitude steps of 250 m from 0 to 9000 m; and finally recalculated by adjusting reference values to equation (1).

$$\vec{\tau}_h = \begin{pmatrix} \tau_B \\ \tau_G \\ \tau_R \\ \tau_{NIR} \\ \tau_{SWIR1} \\ \tau_{SWIR2} \end{pmatrix} = \begin{pmatrix} 0.524225166047 & -0.000171924013 & 2.4651\cdot10^{-8} & -1.25\cdot10^{-12} \\ 0.424690785121 & -0.000142127493 & 2.1028\cdot10^{-8} & -1.08\cdot10^{-12} \\ 0.329870334052 & -0.000117419948 & 1.7625\cdot10^{-8} & -0.91\cdot10^{-12} \\ 0.240047724024 & -0.000096115185 & 1.4375\cdot10^{-8} & -0.73\cdot10^{-12} \\ 0.127035444124 & -0.000048971938 & 0.7141\cdot10^{-8} & -0.36\cdot10^{-12} \\ 0.103740066427 & -0.000035915172 & 0.5254\cdot10^{-8} & -0.27\cdot10^{-12} \end{pmatrix} \cdot \begin{pmatrix} 1 \\ h \\ h^2 \\ h^3 \end{pmatrix} + \begin{pmatrix} c_B \\ c_G \\ c_R \\ c_{NIR} \\ c_{SWIR1} \\ c_{SWIR2} \end{pmatrix} \quad (4)$$

$$\vec{\tau}_h = \begin{pmatrix} \tau_G \\ \tau_R \\ \tau_{NIR1} \\ \tau_{NIR2} \end{pmatrix} = \begin{pmatrix} 0.444322570590 & -0.000148839670 & 2.1868\cdot10^{-8} & -1.1210\cdot10^{-12} \\ 0.338446985809 & -0.000119629348 & 1.7951\cdot10^{-8} & -0.9266\cdot10^{-12} \\ 0.308032298646 & -0.000111491404 & 1.6392\cdot10^{-8} & -0.8353\cdot10^{-12} \\ 0.347199426572 & -0.000123280911 & 1.6401\cdot10^{-8} & -0.7680\cdot10^{-12} \end{pmatrix} \cdot \begin{pmatrix} 1 \\ h \\ h^2 \\ h^3 \end{pmatrix} + \begin{pmatrix} c_G \\ c_R \\ c_{NIR1} \\ c_{NIR2} \end{pmatrix} \quad (5)$$

where τ_h is the atmospheric optical depth for each spectral band, h is the elevation and c is a small additive corrector that is calculated for each band in order to adjust a specific image (date) according to its PIA reflectance values.

Moreover, values of L_a and τ_0 obtained from PIAs are checked with a possible value range. τ_0 thresholds are obtained by the MODTRAN simulations previously explained and L_a thresholds are based on previous experience from several manual corrections under different atmospheric conditions. Therefore, when a value of L_a or $\tau_{0,}$ is over these ranges (Table 1) a specific PIA is not used in a particular correction. This occurs when, on a particular date, the reflectance value is too far (tolerance parameter on Table 1) from the reference value in that PIA (for example because this PIA is under a cloud on the specific date of the image to be corrected).

		B	**R**	**G**	**NIR**		**SWIR1**	**SWIR2**
τ_0 inf	TM	0.265	0.212	0.155		0.097	0.053	0.049
	MSS		0.221	0.160	0.141	0.158		
τ_0 sup	TM	0.600	0.433	0.337		0.250	0.150	0.105
	MSS		0.453	0.346	0.314	0.350		
La inf.		6.93	0.0	2.83		0.0	0.0	0.0
La sup.		34.04	16.0	14.11		6.65	0.13	0.07
tolerance reflectance		0.021	0.015	0.014		0.018	0.028	0.019

Table 1: Range of possible values for atmospheric optical depth and L_a thresholds in digital number units.

It is important to note that the proposed model also includes a topographic correction to account for differences in illumination conditions (solar position at the moment of the image acquisition with respect to surface slope, aspect and elevation) and produces similar reflectance responses for similar terrain features (Vanonckelen et al., 2013), which makes it possible to calculate reflectance in high relief areas accurately (Hantson and Chuvieco, 2011).

Finally, the model also generates different quality indicators, embedded in the image metadata, as well as output quality masks, such as cast shadows.

2. Model application

The model was applied regionally in a heterogeneous area in terms of topography and land cover (see section 3.1. “Application site”). A total of 14 Landsat scenes from

different sensors, processing types and cloud cover were selected, and a DEM was used to run the model (see section 3.2. “Remote sensing imagery and ancillary data”). PIA were generated through a filtering process of 10-year time-series of the Terra-MODIS daily reflectance product (3.3 “Pseudoinvariant area (PIA) generation through MODIS imagery”). Finally, the model performance was evaluated by comparing image classification and spectral signatures (see 3.4 “Model evaluation: spectral signature and image classification”).

2.1. Application site

The 20000 km^2 study area was located in Catalonia, in the northeast of the Iberian Peninsula (Fig. 2), matching the 197-031 path and row in the Worldwide Reference System-2 (WRS-2) used to distribute Landsat 4-7 full frames. The centre of the scene is approximately 2°40'E 41°40'N. The illumination conditions throughout the year are quite different due to latitude (solar elevation angles range from 20.5° to 61.4° at the time of the satellite pass). The area alternates between mountains and plains with a mean elevation of around 700 m a.s.l, but ranging from 0 to 3000 m a.s.l. The rugged surface terrain makes the zone interesting for considering topographic effects on radiometric correction methods. The study area is also composed of different land cover types in a heterogeneous landscape, such as crops, perennial and deciduous woods, shrub areas, urban areas and inland waters, which also makes the site interesting for classification and spectral signature purposes.

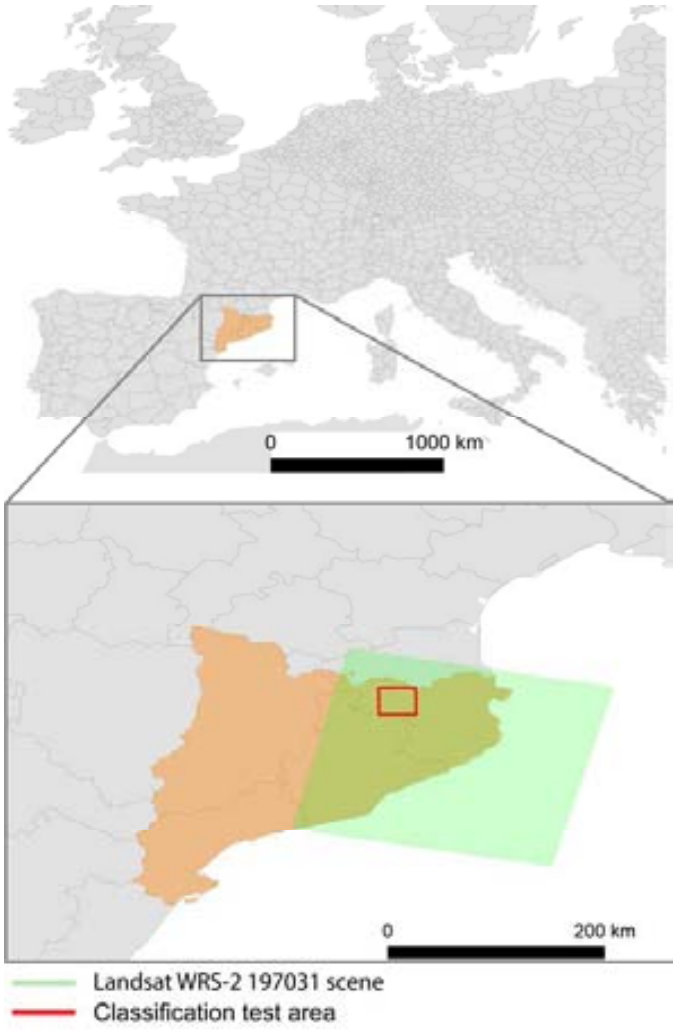


Fig. 2: Study area and model evaluation area (red rectangle).

2.2. Remote sensing imagery and ancillary data

The Radiometric correction model was run using 14 Landsat images, 1 from Landsat-4 TM, 1 from Landsat-5 MSS, 9 from Landsat-5 TM and 3 from Landsat-7 ETM+, all from the 197-031 path and row (see green area in Fig. 2). The images to evaluate the performance of the radiometric correction model were selected in order to consider a wide range of situations. Therefore, selection was based on cloud cover (up to 60%), processing type, format distribution, sensor, spectral signature and classification and image metadata (column “Main role” on Table 2). Image metadata, such as scale and offset parameters to convert from DN to radiance, as well as date/time to compute illumination conditions, are especially important for automating the radiometric process. However, correct metadata treatment is necessary due to the differences in metadata formats between different image distributors and different software capabilities (Pons 2000, Zabala et al., 2002, Pesquer et al., 2012). Landsat images distributed by the

United States Geological Survey (USGS) were downloaded from the EarthExplorer website (http://earthexplorer.usgs.gov), and European Space Agency (ESA) images were available from a previous project (Pons et al., 2012) and completed with an ESA image request (research project 10837). Landsat CDR products, used in the results validation section, were also downloaded from the EarthExplorer website.

Date (yyyy-mm-dd)	Distributor	Type of processing	Platform	Sensor	Main role
1984-06-03	USGS	LPGS	Landsat 5	MSS	type of processing
1985-07-24	USGS	NLAPS	Landsat 5	TM	type of processing
1991-07-01	USGS	NLAPS	Landsat 4	TM	type of processing
2001-09-30	USGS	LPGS	Landsat 7	ETM+	cloud cover
2003-02-08	ESA	CEOS	Landsat 7	ETM+	type of processing
2004-09-30	ESA	CEOS	Landsat 5	TM	spectral signature and classification
2005-04-26	ESA	CEOS	Landsat 5	TM	spectral signature and classification
2005-05-28	ESA	CEOS	Landsat 5	TM	spectral signature and classification
2005-06-29	ESA	CEOS	Landsat 5	TM	spectral signature and classification
2008-06-13	USGS	LPGS	Landsat 7	ETM+	cloud clover + SLC-off
2011-03-10	USGS	LPGS	Landsat 5	TM	classification
2011-04-11	USGS	LPGS	Landsat 5	TM	classification
2011-05-29	USGS	LPGS	Landsat 5	TM	classification
2011-10-04	USGS	LPGS	Landsat 5	TM	classification

Table 2: List of images used to evaluate the radiometric correction.

In order to generate the PIA, a 10-year time-series, from 2002 to 2011, of the Terra-MODIS daily surface reflectance product (MOD09GA) was downloaded from the NASA Earth Observing System Data and Information System (http://reverb.echo.nasa.gov/reverb/). Terra-MODIS reflectance bands similar to Landsat (see Table 3) at 500 m spatial resolution were processed.

Band	Landsat (MSS)	MSS spectral	TM spectral	ETM+ spectral	MODIS band	MODIS spectral

	TM/ETM+ band number	range (nm) (Landsat 4-5)	range (nm)	range (nm)	number	range (nm)
B	1		450-520	450-515	3	459-479
G	(1)2	497-607	520-600	525-605	4	545-565
R	(2)3	603-696	630-690	630-690	1	620-670
NIR	(3)4	701-813	760-900	750-900	2	841-876
	(4)	808-1023				
SWIR 1	5		1550-1750	1550-1750	6	1628-1652
SWIR 2	7		2080-2350	2090-2350	7	2105-2155

Table 3: Correspondence of spectral bands between Terra-MODIS (Feng et al., 2013) and Landsat TM, ETM+ and MSS (Chander et al., 2009).

Finally, to run the topographic correction of the model, we used a DEM of 15 m spatial resolution (MET-15**)** from the Institut Cartogràfic de Catalunya (ICC, 2011).

2.3. Pseudoinvariant area (PIA) generation through MODIS imagery

PIA were generated through the Terra-MODIS surface reflectance product (MOD09GA) at 500 m spatial resolution. This is a very feasible product (Justice et al. 2002), widely used and referenced in the literature for research and applications (*e.g.*, Maier, 2010; Yi et al., 2006). In addition, it has some advantages that make it suitable for generating PIA for the Landsat radiometric correction, such as having a similar spectral configuration (see Table 3) and a similar image acquisition time as the Landsat platforms, minimizing differences in atmospheric and illumination conditions (Feng et al., 2012).

In order to ensure the highest PIA quality, a selection of available MODIS surface reflectance images was made by applying the methodology developed in Pesquer et al. (2013a) based on two criteria: quality masks and reasonable acquisition geometry, and a geostatistical spatial pattern analysis using variograms. In addition, a new, third criterion based on illumination conditions was included in the present work. Indeed, in the original methodology, image pixels were selected according to quality masks (when snow, fire or cloud flags were not present the product was considered of ideal quality) and images for which the sensor zenith angle was higher than 35° at the area of

interest were excluded. The second criterion was used to detect statistical image anomalies through a spatial pattern model obtained from variogram analysis; previous works demonstrated the potential uses of geostatistical tools for analysing spatial patterns (Garrigues et al. 2007; Wallace et al. 2000) as well as for image quality assessment (Pesquer et al. 2013b). In the third, new criterion, illumination conditions were taken into account through a DEM (Wilson and Gallant 2000; Veraverbeke et al. 2010) in order to avoid both cast shadow pixels and pixels under an incidence angle higher than 70° (Proy et al., 1989), considered to be not reliable because they do not usually show Lambertian behaviour for most surfaces; nevertheless, the exigency of being pseudoinvariant causes that selected PIA are mostly located in non rugged areas, so converging both illumination correction protocols and making PIA still more comparable to Landsat imagery. After these three criteria were applied to MODIS imagery, a mean image and a standard deviation image were computed for the 10-year period.

A PIA should show almost constant reflectance values for long time periods. In this study we selected PIA by choosing those pixels that have low standard deviation reflectance values on the image of the standard deviation of the 10-year MODIS series. According to Feng et al. (2013) these thresholds should not be the same for all MODIS bands, and a new set of thresholds has been defined for each band (Table 4).

	B	**R**	**G**	**NIR**	**SWIR1**	**SWIR2**
standard deviation	256	198	186	283	329	245

Table 4: Maximum standard deviation (in 10^{-4} reflectance units, typical for MODIS reflectance products) to be considered as a pseudoinvariant pixel.

2.4. Model evaluation: spectral signatures and image classification

In order to evaluate the performance of the automatic radiometric correction we produced a land cover map of the natural areas, created by classification, and spectral signatures were extracted from the radiometrically corrected Landsat imagery. In the classification, based on the methodology proposed by Serra et al. (2003) and previously applied to other areas (Moré et al., 2007; Serra et al., 2009; Zabala and Pons, 2011; Pons et al., 2012), the results were compared to the classification obtained using imagery corrected by the previous method (non-automatic, and where the atmospheric optical depth is a constant) and to the classification obtained using the

USGS product (when it exists). Two classifications using two sets of images (2011-03-10, 2011-04-11, 2011-05-29, 2011-10-04 and 2004-09-30, 2005-04-26, 2005-05-28, 2005-06-29) were produced. Input variables were Landsat solar bands and NDVI and greenness and wetness Tasseled Cap components (Kauth and Thomas 1976) for each date, using a total of 36 images for each classification. The accuracy was tested through independent test areas (Campbell 1996). Classifications were mainly focused on natural vegetation categories (legend in Fig. 7) in a subregion of the study area of 25 km x 25 km, located in a mountain region, heterogeneous in terms of covers and relief (see red rectangle in Fig. 2).

In addition to image classification, spectral signatures were extracted from Landsat imagery (see Table 2) and compared to the MODIS reflectance product to check for signature coherence. A total of 14 sites with a mean area of 63 ha and at least the area of the MODIS pixel size (500 x 500 m2) corresponding to 4 representative land cover types (Aleppo pine, holm oak tree, Scots pine and urban) were selected using the Land Cover Map of Catalonia (MCSC3, 2005). It was ensured that these areas did not belong to any PIA used to run the radiometric correction.

We considered to also check results through the Landsat–MODIS Consistency Checking System (LMCCS, Feng et al. 2012), but we could not adopt it easily because we used images in different formats and from different processing chains.

3. Results and discussion

3.1. Metadata and image processing type

A total of 14 images from several Landsat platforms (Landsat-4, Landsat-5 and Landsat-7) and sensors (MSS, TM and ETM+) using USGS and ESA file formats and processing chains were automatically corrected (see Table 1). As previously stated, metadata are essential to properly correct remote sensing imagery, especially to avoid errors if an automatic radiometric correction has to be applied. Since the beginning of the Landsat programme, different file formats have been used to distribute Landsat imagery, such as CEOS, GeoTIFF and NDF. These distribution formats depend on the source of acquisition, on the pre-processing level and also on the processing date; this is also true for the metadata files, often changing (Pons et al., 2012; Feng et al., 2012). In addition, these formats are associated with different Landsat image processing systems, particularly to NLAPS and LPGS in the USGS case (Cristóbal et al., 2009). Although the Landsat image distributors continually refine image and metadata distribution, in some of the processed imagery there is a lack of coherence between

image metadata and values in the literature, especially regarding DN to radiances conversion. An important part of this work has been reading multiple formats and ensuring proper metadata handling to apply the most appropriate values according to those in the literature, so avoiding errors caused for these kinds of discrepancies.

3.2. Pseudoinvariant areas (PIA) and atmospheric optical depth

A total of 124 MODIS images (specified in Appendix 1) that fulfilled the PIA filtering criteria were selected from 2002 to 2011. These images cover all months, reproducing a seasonal phenology (see Fig. 3). 444 PIA of 500 m x 500 m were obtained by applying the thresholds from Table 4 to these 124 images. Then, and according to the proposed methodology, equation (1) was applied to each PIA for each Landsat band and date in order to fit parameters L_a and τ_0. In all cases, Table 1 thresholds were guaranteed (in other words, we discarded those Landsat images from dates where fitting L_a and τ_0 implied exceeding the reasonable thresholds).

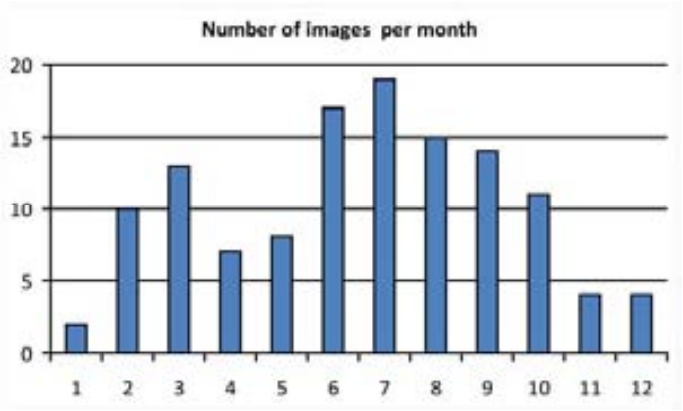


Fig. 3: Monthly distribution of Terra-MODIS best quality images from 2002 to 2011.

Date	B	R	G	NIR	SWIR1*	SWIR2
1984-06-03		42	68	34	107	
1985-07-24	7	36	69	94	213	103
1991-07-01	3	71	77	65	142	112
2001-09-30#	172	11	95	185	209	188
2003-08-02	276	218	211	45	215	153
2004-09-30	268	262	257	261	264	218
2005-04-26	281	277	264	271	276	241
2005-05-28	276	281	262	288	289	241
2005-06-29	249	249	212	270	266	224
2008-06-13#	27	4	28	79	103	89
2011-03-10	281	263	229	211	262	181

2011-04-11	279	245	220	196	254	169
2011-05-29	220	181	159	99	206	136
2011-10-04	290	273	253	250	258	193

Table 5: Number of PIA used per date and Landsat band. B: blue, R: red, G: green, NIR: near infrared, SWIR*: short wave infrared for TM and ETM+, and NIR2 for MSS.

The number of PIA used in each radiometric correction is different depending on the band and date (see Table 5) mainly because each image has different atmospheric conditions. From the image evaluation set, 12 images were successfully corrected with a default tolerance reflectance. However, in two ETM+ images (2001-09-30 and 2008-06-13 flagged with # in Table 5 and Fig. 4), it was necessary to increase that tolerance. Indeed, the combination of high cloud cover (27 % and 58 % respectively) and SLC-off artefacts meant that the minimum number of PIA for a numerical solution was not obtained; in scenes where PIA are hard to find, a reference-based approach (Gao et al. 2010) can be applied. However, as there are some parts of the image that are not affected by clouds it is still interesting to apply the radiometric correction in some cases, for example in drought analysis, in which long time-series of remote sensing imagery are needed (Domingo et al., 2013). In these cases, tolerance could be augmented. Because the proposed method is automatic, image metadata were updated to keep the user informed about the applied tolerance and, consequently, about the potential radiometric correction quality.

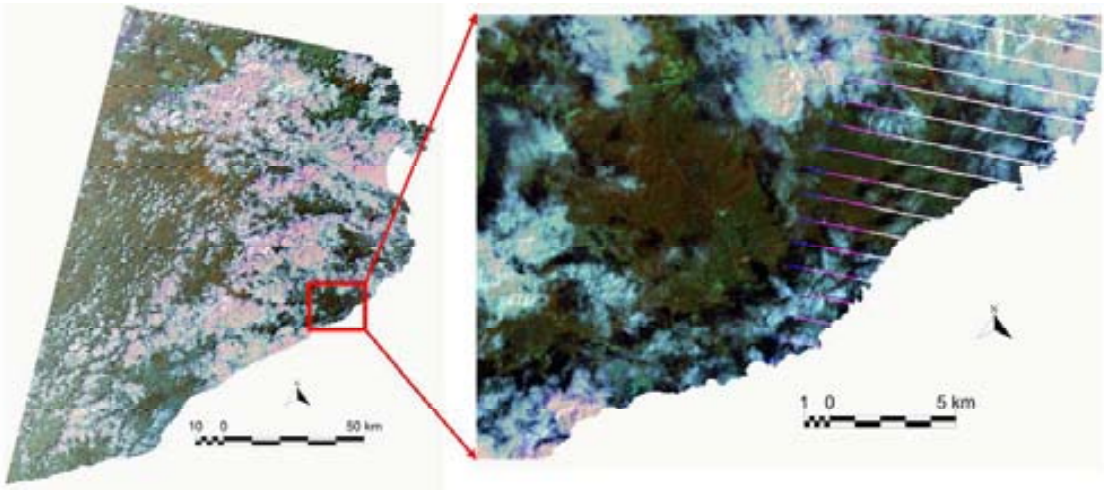


Fig. 4. Landsat-7 ETM+ 4,5,3 RGB composite from 2008-06-13 image after a radiometric correction under very high cloud cover and SLC-off artefacts.

Finally, determining c, the small additive corrector in Eq. 4 based on PIA reference values, allows the continuous model of the atmospheric optical depth to be adapted to

each Landsat band and date, resulting in a more detailed estimation of the atmospheric optical depth (see Fig. 5).

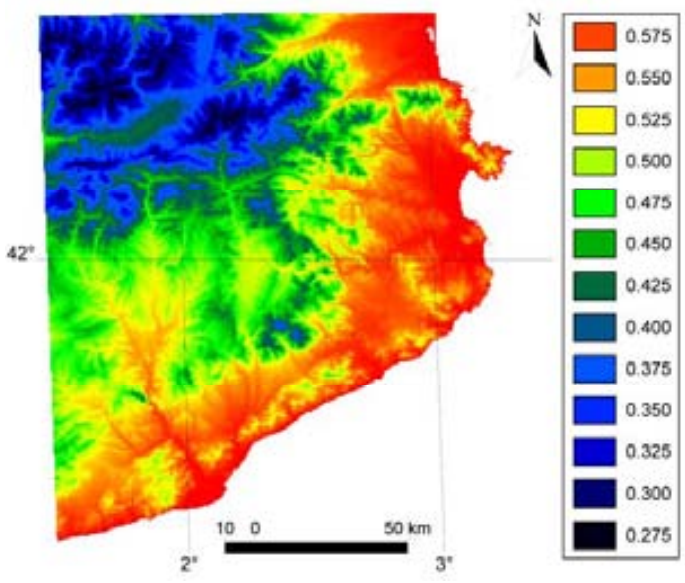


Fig. 5. Atmospheric optical depth computed as a function of elevation for the 2011-05-29 image.

3.3. Classification

Two classifications were generated from three different radiometric correction methods: Pons and Solé (1994) (called “manual” in Tables 6 and 7), the USGS product (called “CDR” in Table 6) and the new automatic method (called “automatic” in Tables 6 and 7). The results in Table 6 and Fig. 6, obtained from an independent set of test areas, show that the accuracy of the maps obtained from the imagery corrected by any of these three methods is similar. Moreover, in the case of the 2004/2005 classification, the automatic methodology slightly improves the manual classification probably due the heterogeneous model of the atmospheric optical depth. This means that the new method is at least as robust as the original manual method, and it makes it possible to correct large volumes of Landsat imagery automatically without any manual supervision once the PIA have been generated.

2011	Automatic	Manual	CDR
Global accuracy	74.70%	76.20%	73.90%

Table 6: Comparison of the global accuracy of the classifications generated by the manual and automatic methods with CDR product for 2011.

2004-2005	Automatic	Manual
Global accuracy	82.30%	80.10%

Table 7: Comparison of the global accuracy of manual and automatic methods for 2004/2005.

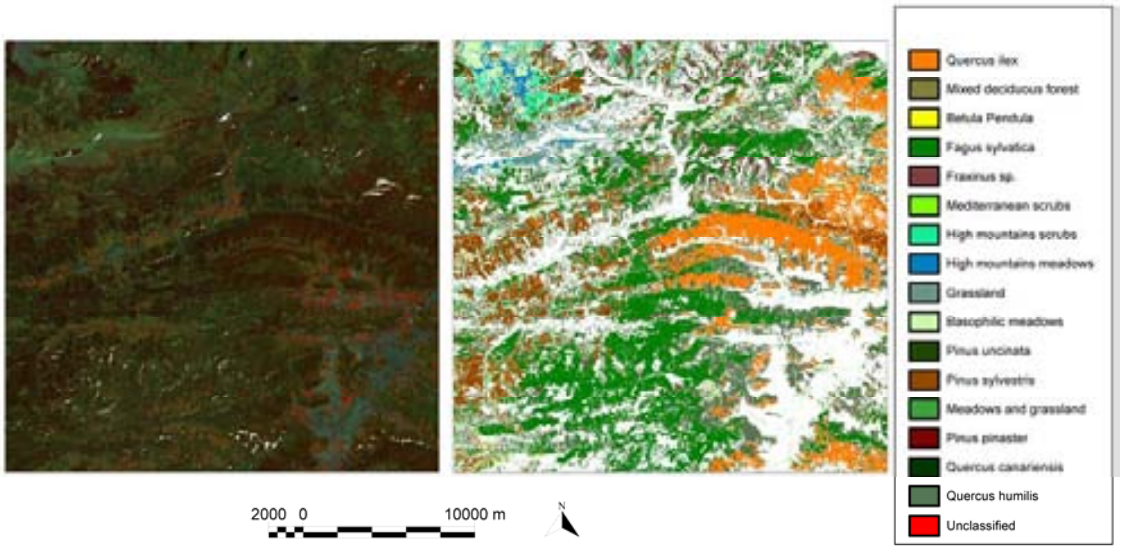


Fig. 6: Left panel: 4-5-3 RGB composition of the reflectance product generated by the automatic method on 26-April-2005. Right panel: result of the natural areas classification using the new reflectance product. "available in color online"

3.4. Spectral signatures

MODIS and Landsat spectral signatures were compared to evaluate the radiometric correction. Spectral signatures of four categories that were representative of the study area (urban, Aleppo pine (*Pinus halepensis*), Scots pine (*Pinus sylvestris*) and holm oak (*Quercus ilex*)) were extracted from several polygons with a minimum area of 28 ha, and a mean reflectance value was computed. As can be seen in Fig. 7, the results are similar to those obtained from MODIS.

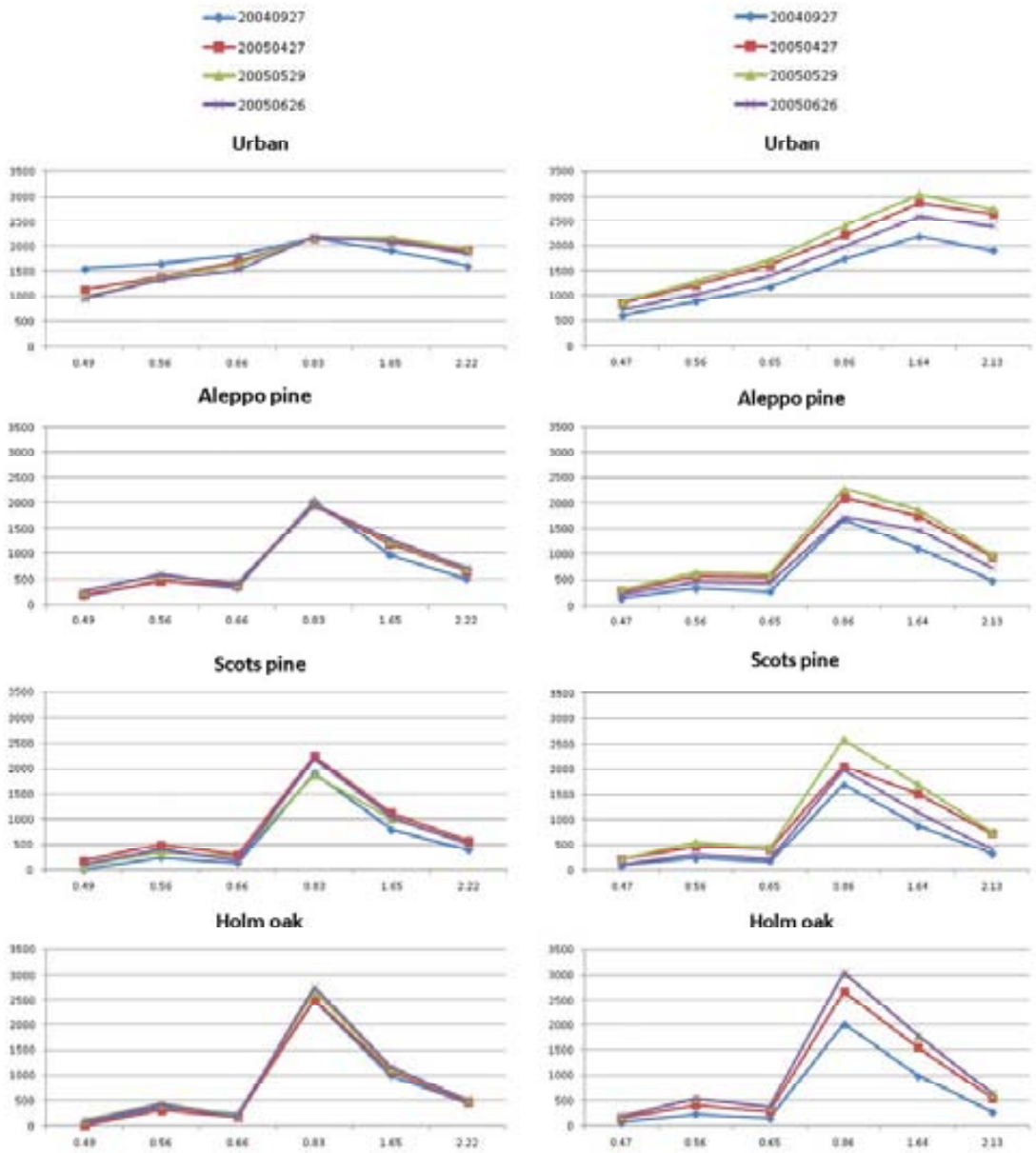


Fig. 7: Spectral signatures of four different categories and four different dates comparing the Landsat automatic radiometric correction (left) and the MODISreflectance product (right).

4. Conclusions

The improved and automated radiometric correction method has been successfully evaluated and ground surface reflectances were obtained for different types of Landsat platforms (Landsat-4, 5 and 7), sensors (MSS, TM and ETM+), formats and processing types (LPGS, NLAPS, CEOS) and even in images with high cloud cover and SCL–off artefacts (Landsat-7). The proposed methodology has demonstrated to be fully automatic, from the selection of the best quality MODIS reference images and the

generation of pseudoinvariant areas (PIA) to the retrieval of ground surface reflectances, without using any auxiliary meteorological or atmospheric products nor requiring dark objects or dense vegetation areas. This is useful in areas with little atmospheric information and allows long time-series of robust imagery to be obtained. It is worth noting that the PIA concept can also be applied to other platforms with similar band configurations, such as SPOT high resolution instruments (HRG, HRVIR or HRV) and the VEGETATION instrument, adapting equations (4) and (5) when required. The evaluation results show good visual agreement between the MODIS and Landsat spectral signatures, and the classification results show that the improved method produces images that are classified with a similar quality to those obtained with the original method; however, now it is possible to correct large volumes of Landsat imagery automatically without any manual supervision once PIA have been generated (also an automatic procedure). Moreover, images can be corrected for a wide variety of situations, providing a larger databank for our area than the CDR product currently does. Classification accuracy is also better with the new method.

As a last comment, we would like to emphasize that this simplified approach is not intended to substitute highly precise radiometric corrections that can be made when additional, more detailed information is available, but rather offers a reasonably good procedure for the new era of long time-series of global remote sensing data. Future work will be focused on the radiometric correction of Landsat-8, SPOT and Sentinel-2 data.

Acknowledgments

This work was partially supported by the Catalan Government under Grant SGR2009-1511, by the Spanish Ministry of Economy and Competitiveness and the European regional development fund (ERDF) under Grant CGL2012-33927. We are very grateful to the USGS policy about Landsat data; the project has also used data provided by the European Space Agency under the ESA research project 10837. Xavier Pons is a recipient of an ICREA Academia Excellence in Research Grant (2011-2015).

References:

Barbosa, P.M., Casterad, M. A., Herrero, J., 1996. Performance of several Landsat 5 Thematic Mapper (TM) image classification methods for crop extent estimates in an irrigation district. International Journal of Remote Sensing 17, 3665-3674

Bariou, R., Lecamus, D., Le Henaff, F., 1986. Corrections Radiometriques, Presses Universitaires de Rennes 2, Rennes.

Barrachina, M., Cristóbal, J., Tulla, A.F., 2010. Los recursos ganaderos en los sistemas extensivos de la montaña pirenaica catalana: aproximación al cálculo de la biomasa herbácea mediante el uso de la Teledetección. Serie Geográfica 16, 35-49.

Bustamante, J., Pacios, F., Díaz-Delgado, R., Aragonés, D., 2009. Predictive models of turbidity and water depth in the Doñana marshes using Landsat TM and ETM+ images. Journal of Environmental Management 90, 2219-2225.

Campbell, J.B., 1996. Introduction to remote sensing, second ed. Taylor and Francis, London

Chander, G., Markham, B.L., Helder D.L., 2009. Summary of current radiometric calibration coefficients for Landsat MSS, TM, ETM+, and EO-1 ALI sensors. Remote Sensing of Environment 113, 893–903.

Chavez, P.S., 1988. An improved dark-object subtraction technique for atmospheric scattering correction of multispectral data. Remote Sensing of Environment 24, 459-479.

Chuvieco, E., Riaño, D., Aguado, I., Cocero, D., 2002a. Estimation of fuel moisture content from multitemporal analysis of Landsat Thematic Mapper reflectance data: Applications in fire danger assessment. International Journal of Remote Sensing 23, 2145-2162.

Chuvieco, E., Martín, M.P., Palacios, A., 2002b. Assessment of different spectral indices in the red-near-infrared spectral domain for burned land discrimination. International Journal of Remote Sensing 23, 5103-5110.

Collado, A.D., Chuvieco, E., Camarasa, A., 2002. Satellite remote sensing analysis to monitor desertification processes in the crop-rangeland boundary of Argentina. Journal of Arid Environments 52, 121-133.

Cristóbal, J., Jiménez-Muñoz, J.C., Sobrino, J.A., Ninyerola, M., Pons, X., 2009 Improvements in Land Surface Temperature Retrieval from the LANDSAT Series Thermal Band Using Water Vapour and Air Temperature. Journal of Geophysical Research: Atmospheres 11.

Cristóbal, J., Poyatos, R., Ninyerola, M., Llorens, P., Pons, X., 2011. Combining remote sensing and GIS climate modelling to estimate daily forest evapotranspiration in a Mediterranean mountain area. Hydrology and Earth System Sciences 15, 1563-1575.

Domingo, C., Cristóbal, J., Ninyerola, M., Pons, X., 2013. MODIS time series analysis as a tool for forest drought detection in Catalonia (NE Iberian Peninsula): integration of remote sensing and climatic variables. Geophysical Research Abstracts 15, EGU2013-10300. EGU General Assembly 2013.

Dozier, J. 1989. Spectral signature of Alpine snow cover from the Landsat Thematic Mapper. Remote Sensing of Environment 28, 9-22.

European Spatial Agency (2005) Landsat ETM/TM CEOS/ESA Products. The internet: http://earth.esa.int/pub/ESA_DOC/Landsat_FAQ.pdf (accessed on 2-1-2014).

Feng, M., Huang, C. , Channan, S., Vermote, E. F., Masek, J. G., Townshend J.R., 2012. Quality assessment of Landsat surface reflectance products using MODIS data. Computers & Geosciences 38, 9-22

Feng, M., Sexton, J.O., Huang C., Masek, J.G., Vermote, E.F., Gao, F., Narasimhan, R., Channan, S., Wolfe, R.E., Townshend J.R., 2013. Global surface reflectance products-from Landsat-Assessment using coincident MODIS observations. Remote Sensing of Environment 134, 276-293.

Franklin S.E., Giles P.T., 1995. Radiometric processing of aerial and satellite remote-sensing imagery. Computers & Geosciences 21, 413-423.

Gao, F., Masek, J., Wolfe, R., Huang, C., 2010, Building consistent medium resolution satellite data set using moderate resolution imaging spectroradiometer products as reference, Journal of Applied Remote Sensing, 4, 043526, doi:10.1117/1.3430002.

García-Millán, V., Sánchez-Azofeifa, G.A., Malvárez, G.-C., Moré, G., Pons, X., Yamanaka-Ocampo, M., 2013. Effects of Topography on the Radiometry of CHRIS/PROBA Images of Successional Stages Within Tropical Dry Forests. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 6, 1584-1595

Garrigues, S., Allard, D., Baret, F., 2007. Using First and Second Order Variograms for Charac-terizing Landscape Spatial Structures from Remote Sensing Imagery. IEEE Transactions on Geoscience and Remote Sensing, 45, 1823-1834.

Goward, S.N., Masek, J.G., 2001. Editorial: Landsat-30 years and counting. Remote Sensing of Environment 78, 1-2.

Hadjimitsis, D.G., Clayton, C.R.I., Retalis, A., 2009. The use of selected pseudo-invariant targets for the application of atmospheric correction in multi-temporal studies using satellite remotely sensed imagery. International Journal of Applied Earth Observation and Geoinformation 11, 192-200.

Hale, S.R., Rock, B.N., 2003. Impact of topographic normalizati on on land-cover classification accuracy Photogrammetric Engineering and Remote Sensing 69, 785-791.

Hantson, S., Chuvieco, E., 2011. Evaluation of different topographic correction methods for Landsat imagery. International Journal of Applied Earth Observation and Geoinformation 13, 691-700.

ICC 2011 Especificacions tècniques Revisió de document 2 - Juny 2011 del Model d'Elevacions del Terreny de Catalunya 15x15 metres (MET-15) Institut Cartogràfic de Catalunya. The internet: http://www.icc.cat/cat/content/download/12339/41446/file/met15v20esp_02ca.pdf (accessed 2-Jan-2014)

Janzen, D.T., Fredeen, A.L., Wheate, R.D., 2006. Radiometric correction techniques and accuracy assessment for Landsat TM data in remote forested regions. Canadian Journal of Remote Sensing, 32, 330-340.

Justice, C.O., Townshend, J.R., Vermote, E.F., Masuoka, E., Wolfe, R.E, Saleous, N., Roy, D.P., Morisette, J.T., 2002. An overview of MODIS Land data processing and product status Remote Sensing of Environment 83, 3-15.

Kauth, R.J., Thomas, G.S., 1976. The tasseled Cap. A Graphic Description of the Spectral-Temporal Development of Agricultural Crops as Seen by LANDSAT. Proceedings of the Symposium on Machine Processing of Remotely Sensed Data, Purdue University of West Lafayette, Indiana, 4B41-4B51.

Kotchenova, S.Y., Vermote, E.F., 2007. Validation of a vector version of the 6S radiative transfer code for atmospheric correction of satellite data. PartII. Homogeneous Lambertian and anisotropic surfaces. AppliedOptics 46(20), 4455-4464. doi:10.1364/AO.46.004455.

Liu, Y.S., Hu, Y.C., Peng, L.Y., 2005. Accurate quantification of grassland cover density in an alpine meadow soil based on remote sensing and GPS. Pedosphere 15, 778-783.

Lopes, D.M., Aranha, J.T., Walford, N., O'Brien J., Lucas N., 2009. Accuracy of remote sensing data versus other sources of information for estimating net primary production in Eucalyptus globulus Labill. and Pinus pinaster Ait. ecosystems in Portugal. Canadian Journal of Remote Sensing 35, 37-53.

Maier, S.W., 2010. Changes in surface reflectance from wildfires on the Australian continent measured by MODIS. International Journal of Remote Sensing 31, 3161-3176.

Masek, J., Vermote, E., Saleous, N., Wolfe, R., Hall, F., Huemmrich, K.F., Gao, F., Kutler, J., Lim T., 2006. A Landsat surface reflectance dataset for NorthAmerica, 1990–2000. IEEE Geoscience and Remote Sensing Letters 3, 68-72

MCSC3, 2005. Land Cover Map of Catalonia, 3rd edition version 2 (2005-2007) The Internet: http://www.creaf.uab.es/mcsc/usa/index.htm (accessed on 2-1-2014).

MODTRAN, 2012. MODerate resolution atmospheric TRANsmission: Narrow band model atmospheric radiative transfer code, v. 5. The Internet: http://www.modtran5.com (accessed on 2-1-2014).

Moré, G., Serra. P., Pons. X., 2007. Improvements on Classification by Tolerating NoData Values - Application to a Hybrid Classifier to Discriminate Mediterranean Vegetation with a Detailed Legend Using Multitemporal Series of Images. IEEE International Geoscience and Remote Sensing Symposium, IGARSS; Denver, CO; United States, 4241201 1:192-195.

Moré, G., Serra. P., Pons. X., 2011. Multitemporal flooding dynamics of rice fields by means of discriminant analysis of radiometrically corrected remote sensing imagery. International Journal of Remote Sensing 32 1983-2011.

Newton, A.C., Hill, R.A., Echeverría, C., Golicher, D., Rey Benayas, J.M., Cayuela, L., Hinsley, S.A., 2009. Remote sensing and the future of landscape ecology. Progress in Physical Geography 33, 528-546.

Nuarsa, I.W., Nishio, F., Hongo, C., 2010. Development of the empirical model for rice field distribution mapping using multi-temporal Landsat ETM+ data: case study in Bali

Indonesia. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Science, XXXVIII, part 8, Kyoto Japan.

Oliveras, I., Gracia, M, Moré, G., Retana, J., 2009. Factors influencing the pattern of fire severities in a large wildfire under extreme meteorological conditions in the Mediterranean basin. International Journal of Wildland Fire 18, 755–764

Pérez-Cabello, F., Ibarra, P, Echeverría, M.T., de la Riva, J., 2010, Post-fire land degradation of Pinus sylvestris L. woodlands after 14 years. Land Degradation and Development 21, 145-160.

Pesquer, L., Prat, E., Díaz-Delgado, R., Masó, J., Bustamante, J., Pons, X., 2012. Automatic modelling and continuous map generation from georeferenced species census data in an interoperable GIS environment. Proceedings of International Environmental Modelling and Software Society, Leizpig, Germany. ISBN: 978-88-9035-742-8.

Pesquer, L., Domingo, C., Pons, X., 2013a. A Geostatistical Approach for Selecting the Highest Quality MODIS Daily Images. Springer Lecture Notes in Computer Science Series, 7887 LNCS, 608-615.

Pesquer, L., Pons, X., Cortés, A., Serral, I., 2013b. Spatial pattern alterations of JPEG2000 lossy compression in remote sensing images. Massive variogram analysis in High Performance Computing. Journal of Applied Remote Sensing 73595.

Pons, X., Solé-Sugrañes, L., 1994. A Simple Radiometric Correction Model to Improve Automatic Mapping of Vegetation from Multispectral Satellite Data. Remote Sensing of Environment 48, 191-204.

Pons X., 2000. MiraMon. Geographical Information System and Remote Sensing Software, Centre for Ecological Research and Forestry Applications, CREAF The Internet: http://www.creaf.uab.cat/MiraMon (accessed on 2-1-2014)

Pons, X., Arcalís, A., 2012. Diccionari terminològic de Teledetecció. Enciclopèdia Catalana and Institut Cartogràfic de Catalunya, Barcelona. ISBN: 978-84-412-2249-6.

Pons, X., Cristóbal, J., González, O., Riverola, A., Serra, P., Cea, C., Domingo, C., Díaz, P., Monterde, M., Velasco, E., 2012. Ten years of Local Water Resource Management: Integrating Satellite Remote Sensing and Geographical Information Systems. European Journal of Remote Sensing 45, 317-332.

Pons, X., Ninyerola, M., Cea, C., González-Guerrero, Ò., Serra, P., Zabala, A., Pesquer, L., Serral, I., Masó, J., Domingo, C., Serra, J.M., Cristóbal, J., Hain, C.R., Anderson, M.C., 2014. Preparing for global land cover & climate change mapping at detailed resolution. The design of a massive database from long time series of Landsat land cover products and in situ climate data. Global Vegetation Monitoring and Modeling symposium, GV2M. Avignon. Paper S2.16.

Potapov, P., Hansen, M.C., Stehman, S.V., Loveland, T.R., Pittman, K., 2008. Combining MODIS and Landsat imagery to estimate and map boreal forest cover loss. Remote Sensing of Environment 112, 3708-3719.

Proy, C., Tanrd, D., Deschamps, P.Y., 1989. Evaluation of Topographic Effects in Remotely Sensed Data Remote Sensing of Environment 30, 21-32.

Rabus, B., Eineder, M., Roth, A., Bamler, R., 2003. The shuttle radar topography mission- a new class of digital elevation models acquired by spaceborne radar. ISPRS Journal of Phototgrammetry & Remote Sensing 57, 241-262.

Riaño, D., Chuvieco E., Salas J., Aguado I., 2003. Assessment of Different Topographic Corrections in Landsat-TM Data for Mapping Vegetation Types. IEEE Transactions on Geoscience and Remote Sensing 41, 1056-1061.

Román - Cuesta, R.M., Retana, J., Gracia, M., Rodríguez, R., 2005. A quantitative comparison of methods for classifying burned areas with LISS - III imagery. International Journal of Remote Sensing 26, 1979-2003.

Richards, J. , Jia. X., 2005. Remote Sensing Digital Image Analysis: An Introduction, fourth ed., Springler-Verlag, Berlin.

Roy, D., Borak, J., Devadiga, S., Wolfe, R., Zheng, M., Descloitres, J., 2002. The MODIS land product quality assessment approach. Remote Sensing of Environment 83, 62-76.

Salvador, R., Pons. X., Baulies, X., 1997. Análisis de imágenes multiespectrales aerotransportadas para estimar variables estructurales de bosques mediterráneos de Quercus ilex L. Orsis 12, 127-139.

Sánchez, N., Martínez-Fernández, J., Calera, A., Torres, E., Pérez-Gutiérrez, C., 2010. Combining remote sensing and in situ soil moisture data for the application and

validation of a distributed water balance model (HIDROMORE). Agricultural Water Management 98, 69–78.

Schroeder, T.A., Cohen, W.B., Song, C., Canty, M.J., Yang, Z., 2006. Radiometric correction of multi-temporal Landsat data for characterization of early successional forest patterns in western Oregon. Remote Sensing of Environment 103, 16-26.

Serra, P., Pons. X., Saurí. D., 2003. Post-classification change detection with data from different sensors: some accuracy considerations. International Journal of Remote Sensing 24, 3311-3340.

Serra, P., Pons. X., 2008. Monitoring farmers' decisions on Mediterranean irrigated crops using satellite image time series. International Journal of Remote Sensing 29, 2293-2316.

Serra P, Moré G, Pons X, 2009. Thematic accuracy consequences in cadaster land-cover enrichment from a pixel and from a polygon perspective. Photogrammetric Engineering and Remote Sensing 75 (12): 1441-1449.

Serra, P., Queralt, E., Pin, C., González, O., 2012. Remote Sensig Methods Applied to Quantify the Water Balance in an Old Irrigation Community. IEEE Geoscience and Remote Sensing Symposium, IGARSS-2012, paper THP.P.759

Slater, J.A., Heady, B., Kroenung, G., Curtis, W., Haase J., Hoegemann, D., Shockley, C., Tracy, K., 2011. Global Assessment of the New ASTER Global Digital Elevation Model. Photogrammetric Engineering & Remote Sensing 77, 335-350.

Song, C., Woodcock, C.E., Seto, K.C., Lenney, M.P., Macomber,S.A., 2001. Classification and change detection using Landsat TM data when and how to correct atmospheric effects? Remote Sensing of Environment 75, 230-244.

Themistocleous, K., Hadjimitsis, D.G., Retalis, A., Chrysoulakis, N., 2012. Development of a new image based atmospheric correction algorithm for aerosol optical thickness retrieval using the darkest pixel method. Journal of Applied Remote Sensing 6, 063538.

U.S. Geological Survey. Product guide: Landsat climate data record (CDR). Surface reflectance. Department of the Interior U.S. Geological Survey. Version 3.4, December 2013. The Internet: http://landsat.usgs.gov/documents/cdr_sr_product_guide.pdf (accessed on 2-Jan-2014).

Vanonckelen, S., Lhermitte, S., Van Rompaey, A., 2013. The effect of atmospheric and topographic correction methods on land cover classification accuracy International Journal of Applied Earth Observation and Geoinformation 24, 9-21.

Vázquez, A., 2008, Structural attributes of three forest types in central Spain and Landsat ETM+ information evaluated with redundancy analysis. International Journal of Remote Sensing 29(19): 5657-5676

Veraverbeke, S., Verstraeten, W.W., Lhermitte, S., Goossens, R., 2010. Illumination effects on the differenced Normalized Burn Ratios optimality for assessing fire severity. International Journal of Applied Earth Observation and Geoinformation 12, 60-70.

Vermote, E.F., Vogelmann, J., Wulder, M.A., Wynne, R., 2008. Free access to landsat imagery Science 320(5879):1011.

Vermote, E.F., Kotchenova, S.Y., 2008. Atmospheric correction for the monitoring of land surfaces. Journal of Geophysical Research 113(D23), D23S90. doi:10.1029/2007JD009662.

Vicente-Serrano, S.M., Pérez-Cabello, F., Lasanta, T., 2008. Assessment of radiometric correction techniques in analyzing vegetation variability and change using time series of Landsat images Remote Sensing of Environment 112, 3916-3934.

Wallace, C.S.A., Watts, J. M., Yool, S.R., 2000. Characterizing the spatial structure of vegetation communities in the Mojave Desert using geostatistical techniques. Computers & Geosciences 26, 397-410.

Wilson, J.P., Gallant, J.C., 2000. Terrain Analysis. Principles and Applications. John Wiley & Sons, New York.

Yi, Y., Yang, D., Huang, J., Chen, D., 2008. Evaluation of MODIS surface reflectance products for wheat leaf area index (LAI) retrieval. ISPRS Journal of Photogrammetry and Remote Sensing 63, 661-677.

Zabala, A., Pons, X., 2002. Image Metadata: compiled proposal and implementation. Benes T (ed.) Geoinformation for European-wide Integration. Millpress, Rotterdam, 647-652. ISBN: 90-77017-71-2.

Zabala A., Gonzalez-Conejero J., Serra-Sagristà J., Pons X., 2010. JPEG2000 encoding of images with NODATA regions for Remote Sensing applications. Journal of Applied Remote Sensing 041793.

Zabala, A., Pons, X. 2011. Effects of lossy compression on remote sensing image classification of forest areas. International Journal of Applied Earth Observation and Geoinformation 13, 43-51.

Zha, Y., Gao, J., Ni, S., Liu, Y., Jiang J., Wei, Y., 2003. A spectral reflectance-based approach to quantification of grassland cover from Landsat TM imagery. Remote Sensing of Environment 87, 371–375.

Zha, Y., Gao, J., Nia, S., Shena, N., 2005, Temporal filtering of successive MODIS data in monitoring a locust outbreak. International Journal of Remote Sensing 26, 5665-5674

Appendix 1

List of the 124 images (MODIS Terra MOD09GA product) selected for generating pseudoinvariant areas. They are grouped by year.

20020615; 20020622; 20020814; 20020915; 20020926
20030208; 20030312; 20030321; 20030406; 20030614; 20030625; 20030630; 20030711; 20030718; 20030803; 20030810; 20030918; 20031105
20040213; 20040424; 20040618; 20040722; 20040814; 20040922; 20040927;
20050213; 20050319; 20050427; 20050506; 20050525; 20050529; 20050626; 20050716; 20050806; 20050831; 20051223;
20060104; 20060126; 20060207; 20060214; 20060228; 20060302; 20060311; 20060327; 20060331; 20060412; 20060509; 20060519; 20060525; 20060601; 20060608; 20060613; 20060619; 20060624; 20060703; 20060710; 20060721; 20060726; 20060731; 20060804; 20060809; 20060903; 20060910; 20061007; 20061028; 20061030; 20061106; 20061110; 20061212; 20061226; 20061228; 20061231;
20070314; 20070424; 20070508; 20070602; 20070706; 20070715; 20070805; 20070828; 20070901; 20070906; 20071019; 20071102;
20080208; 20080222; 20080302; 20080314; 20080622; 20080701; 20080720; 20080724; 20080731; 20080805; 20080915; 20081005;
20090226; 20090314; 20090319; 20090718; 20090723; 20090817; 20090821; 20090906; 20090929; 20091013;
20100520; 20100621; 20100705; 20100714; 20100930; 20101007; 20101018;
20110205; 20110307; 20110401; 20110412; 20110622; 20110627; 20110811; 20110823; 20110910; 20111003; 20111010; 20111026.

Annex 2. Referències

Referències de la Introducció, Resum dels resultats i Conclusions finals

Agrawal, D. Bernstein P., Bertino E., Davidson S., Dayal U., Franklin, M., Gehrke, J., Haas, L., Halevy, A., Han, J., Jagadish, H.V., Labrinidis A., Madden, S., Papakonstantinou Y., Patel J.M., Ramakrishnan R., Ross K., Shahabi C., Suciu D.,Vaithyanathan S., Widom, J. (2012) Challenges and Opportunities with Big Data – A community white paper developed by leading researchers across the United States. Computing Research Association.

Balaguer-Beser A., Ruiz L.A., Hermosilla T., Recio J.A. (2013) Using semivariogram indices to analyse heterogeneity in spatial patterns in remotely sensed images, *Computers and Geosciences* 50,115-127.

Balaji P., Buntinas D., Goodell D., Gropp W., Hoefler T., Kumar S., Lusk E., Thakur R., Träff J.L. (2011) MPI on millions of cores. *Parallel Processing Letters* 21(1), 45–60.

Berry B.J.L., Marble D. (1968) *Spatial Analysis: a reader in statistical geography*. Englewood Cliffs, NJ: Prentice Hall.

Bosque-Sendra J. (2005) Espacio geográfico y Ciencias sociales. Nuevas propuestas para el estudio del territorio. *Investigaciones regionales* 6. 203-224.

Burrough P.A., McDonnell R.A., 1998. *Principles of Geographical Information Systems*. Oxford University Press, 333 pp.

Committee on Strategic Directions for the Geographical Sciences in the Next Decade; National Research Council (2010). *Understanding the Changing Planet: Strategic Directions for the Geographical Sciences*. National Academies Press, 500 Fifth Street, N.W., Lockbox 285, Washington, DC.

Cressie N.A.C.(1993) *Statistics for Spatial Data*. Wiley Series in Probability and Mathematica Statistics Johm Wiley & Sons New York, pp. 900.

Curran P.J., Atkinson P. M. (1998) Geostatistics and remote sensing. *Progress in Physical Geography*, 22(1), 61-78.

Curran P.J. (2001). Remote sensing: Using the spatial domain. *Environmental and Ecological Statistics* 8(4), 331-344.

Chica-Olmo M., Abarca-Hernandez F (1998) Radiometric coregionalization of Landsat TM and SPOT HRV images. *International Journal of Remote Sensing* 19(5), 997-1005

Chuvieco E., Bosque Sendra J., Pons X., Conesa C., Santos-Preciado J.M, Gutiérrez Puebla J., Salado M.J., Martín M.P., Riva, J. de la, (2005). ¿Son las Tecnologías de la Información Geográfica (TIG) parte del núcleo de la Geografía? *Boletín de la AGE* 40: 35-55 .

Fotheringham A.S., Brunsdon C,. Charlton M. (2000) *Quantitative Geography: Perspectives on Spatial Data Analysis*. London; Thousand Oaks, Calif., Sage Publications.

Goodchild M.F. (1992): Geographical information science. *International Journal of Geographical Information Systems* 6(1), 31-45.

Goovaerts P. (1997) *Geostatistics for Natural Resources Evaluation.* Oxford University Press, New York.

Griffith D.A. (2012) Spatial statistics: A quantitative geographer's perspective. *Spatial Statistics*, 1 3-15.

Haining R.P., Kerry R., Oliver M.A. (2010) Geography, Spatial Data Analysis, and Geostatistics: An Overview. *Geographical Analysis* 42, 7-31

Hiemstra P.H., Pebesma E.J., Twenhofel C.J.W., Heuvelink G B.M. (2007). Toward an automatic real-time mapping system for radiation hazards. In: Klien, E. (Ed.), *GI-Days conference*. Institut fur Geoinformatik, Munster, Germany, p. 6.

Hengl T.(2009) *A Practical Guide to Geostatistical Mapping of Environmental Variables*. *2nd edition*. EUR 22904 EN Scientific and Technical Research series report published by Office for Official Publications of the European Communities, Luxembourg.

Irish R.R. (2000) Landsat 7 science data user's handbook, Report 430-15-01-003-0 , National Aeronautics and Space Administration

Jian X, Olea R.A,. Yu Y. (1996) Semivariogram modeling by weighted least squares. *Computers & Geosciences* 22 (4), 387-397

King R.S. (2013) The top 10 programming languages I*EEE Spectrum* 48 (10).

Kitanidis P.K. (1997) *Introduction to geostatistics: applications to hydrogeology*., Cambridge University Press. 1997, p. 249

Labrinidis A., Jagadish H.V. (2012) Challenges and Opportunities with Big Data. *Proceedings of the VLDB Endowment* 5(12) 2032-2033.

Laney D. (2001) 3D Data Management: Controlling Data Volume, Velocity and Variety. *Gartner*.

Lynch C. (2008) Big data: how do your data grow? *Nature* 455 (7209), 28e29.

Madden S. (2012) From databases to big data. *IEEE Internet Computing* 16 (3), article number 6188576, 4-6.

Nunes J. (2014) Web ICGC 1.0. Institut Cartogràfic i Geològic de Catalunya http://www.icc.cat/cat/Home-ICC/Mapes-escolars-i-divulgacio/Diccionaris/Analisi-espacial (accessible el 23-04-2014).

Oliver M.A., Shine J.A., Slocum K.R. (2005). Using the variogram to explore imagery of two different spatial resolutions. *International Journal of Remote Sensing* 26, 3225–3240.

Pons X., Solé-Sugrañes L. (1994) A Simple Radiometric Correction Model to Improve Automatic Mapping of Vegetation from Multispectral Satellite Data. *Remote Sensing of Environment* 48, 191-204.

Pons X., Arcalís A. (2012) *Diccionari terminològic de Teledetecció*. Enciclopèdia Catalana i Institut Cartogràfic de Catalunya, Barcelona. 597 p.

Poon J.P.H. (2003). Quantitative methods: Producing quantitative methods narratives. *Progress in Human Geography* 27(6): 753-62.

Reitsma F. (2013) Revisiting the Is GIScience a science? debate (or quite possibly scientific gerrymandering) *International Journal of Geographical Information Science* 27:2, 211-221.

Ringrose S., Venderpost C., Matheson W. (1996). The use of integrated remotely sensed and GIS data to determine causes of vegetation cover change in southern Botswana. *Applied Geography* 16, 225-242.

Rodgers S.E., Oliver M.A. (2007) A Geostatistical Analysis of Soil, Vegetation, and Image Data Characterizing Land Surface Variation. *Geographical Analysis* 39, 195-216.

Slater J.A., Heady B., Kroenung G., Curtis W., Haase J., Hoegemann D., Shockley C., Tracy K., (2011). Global Assessment of the New ASTER Global Digital Elevation Model. *Photogrammetric Engineering & Remote Sensing* 77, 335-350.

Singh S., Singh N. (2012) Big Data Analytics. *International Conference on Communication, Information & Computing Technology* (ICCICT), Mumbai, India

Turner M.D. (2003) Methodological Reflections on the Use of Remote Sensing and Geographic Information Science in Human Ecological Research. *Human Ecology* 31(2), 255-279.

Unwin D.J. (1996) GIS, spatial analysis and spatial statistics *Progress in Human Geography* 20, 540-551.

Walker D.W., Dongarra J.J. (1996) MPI: A standard message passing interface . *Supercomputer* 12(1), 56-68.

Woodcock CE, Allen R, Anderson M, Belward A, Bindschadler R, Cohen W, Gao F, Goward SN, Helder D, Helmer E, Nemani R, Oreopoulos L, Schott J, Thenkabail PS, Vermote EF, Vogelmann J, Wulder MA, Wynne R. (2008). Free access to Landsat imagery. *Science* 320, 1011.

Wright D.J., Goodchild M.F. Proctor J.D. (1997) Demystifying the persistent ambiguity of GIS as 'tool' versus 'science. *Annals of the Association of American Geographers* 87(2), 346-362.

Yang X., Blower J.D., Bastin L., Lush V., Zabala A., Masó J., Cornford D., Díaz P., Lumsden J. (2013) An Integrated View of Data Quality in Earth Observation. *Philosophical Transactions of the Royal Society* A 371: 20120072.

Zabala A., Pons X. (2002) Image Metadata: compiled proposal and implementation. Benes T (ed.) *Geoinformation for European-wide Integration*. Millpress, Rotterdam, 647-652. ISBN: 90-77017-71-2.

Annex 3. Citacions Capítol 2

Cheng T., (2013) Accelerating universal Kriging interpolation algorithm using CUDA-enabled GPU. *Computers and Geosciences,* 54, 178-183.

De Mesnard (2013) Pollution models and inverse distance weighting: Some critical remarks. *Computers and Geosciences*, 52, 459-469.

Gutiérrez de Ravé E., Jiménez-Hornero F.J., Ariza-Villaverde A.B., Gómez-López J.M. (2014) Using general-purpose computing on graphics processing units (GPGPU) to accelerate the ordinary kriging algorithm. *Computers and Geosciences*, 64, 1-6.

Nowak W., Litvinenko A. (2013) Kriging and Spatial Design Accelerated by Orders of Magnitude: Combining Low-Rank Covariance Approximations with FFT-Techniques. *Mathematical Geosciences*, 45(4) 411-435.

Shi X., Ye F. (2013) Kriging interpolation over heterogeneous computer architectures and systems. *GIScience and Remote Sensing* 50(2) 196-211.

Song X., Liu X., Tang G., Wang Y., Tian J., Dou W. (2012) Parallel Computing of the Digital Elevation Model and Digital Terrain Analysis. *Geography and Geo-Information Science* 8(4), 208.

Yu X. , Deng W., Xiao K., Zou W. (2013) Visualized reserves estimation based on 3D Kriging method. *Earth Science Frontiers* 20(4) 320-331.