ОСИГУРЯВАНЕ И КОНТРОЛ НА КАЧЕСТВОТО НА ДАННИ ЗА ЗЕМНОТО ПОКРИТИЕ ПОЛУЧЕНИ ОТ СПЪТНИКОВИ ИЗОБРАЖЕНИЯ

Венцеслав Димитров

Институт за космически изследвания и технологии – Българска академия на науките
e-mail: vdimitro@stil.bas.bg

Ключови думи: *качество на данните, тематични данни, оценка на точността, спътникови изображения, класификация*

Резюме: *Качествените данни са предпоставка за приемане на информирани решения и осигуряване на предвидими резултати. При проекти, генериращ данни, е нужно текущо да се прилагат мерки за осигуряване на качеството на данните, а накрая - да се извърши неговата оценка. Докладът се занимава с проблема за осигуряване и контрол на тематичната точност като елемент на качеството на данните за земното покритие на страната, получени чрез автоматизирана класификация на спътникови изображения. След като, в хода на проекта, междинните продукти са били подложени на верификация и подобряване, крайните продукти се валидират преди тяхното окончателно приемане. На оценка е подложен слоят за степента на непроницаемост за България от 2012 г., получен по програмата "Коперник". Тематичната точност се определя чрез стратифицирана случайна извадка и сравняване с референтни данни с по-висока пространствена разделителна способност. Получената грешка от неправилна класификация показва, че е нужно да се подобри методологията за извличане на тематичната информация. Количествените параметри от диаграмата на разсейване на плътностите на непроницаемост също могат да се използват за тази цел.*

QUALITY ASSURANCE AND CONTROL OF LAND COVER DATA EXTRACTED FROM SATELLITE IMAGES

Ventzeslav Dimitrov

Space Research and Technology Institute – Bulgarian Academy of Sciences
e-mail: vdimitro@stil.bas.bg

Keywords: *data quality, accuracy assessment, thematic data, satellite image, classification*

Abstract: *Qualitative data is a prerequisite for informed decisions and a guarantee for predictable results. During projects producing data it is throughout necessary to apply measures to ensure required data quality and also final quality assessment. This paper deals with the problem of assurance and control of the thematic accuracy as an element of the quality of land cover data derived from satellite imagery through automated classification. In the course of the project intermediate products were subject to verification and improvement and the final products are validated for their final adoption. The Degree of Imperviousness 2012 layer for Bulgaria is examined which was created during the Copernicus programme. The thematic accuracy is evaluated by stratified random sampling and comparison with reference data of higher spatial resolution. The obtained commission error shows that the thematic information extraction methodology needs some improvements. Quantitative parameters received from impervious densities scatter-plot can be used in the same direction as well.*

Въведение

Качество е съвкупност от свойства и характеристики на даден продукт или услуга, които удовлетворяват съществуващи или предполагаеми потребности на потребителите. Качество на данните е характеристика на данните, показваща степента на възможност за техния анализ [1]. Основни критерии на качеството на данните се явяват пълнотата, точността, логическата

съгласуваност и актуалността. Оценката на качеството на данните е необходима стъпка във всеки аналитичен проект, защото, при недостатъчно високо качество, приетите решения ще бъдат неточни и недостоверни.

Пространствените данни представят различни аспекти от реалния пространсвен свят и тяхното качество може да се разглежда като различие между тази представа и реалността [2]. Няколко от стандартите на серията ISO 19100 са посветени на качеството на географската информация, а според ISO 19113 позиционната и тематичната точност са два от петте елемента на качеството на данните. В редицата значими европейски проекти за космическо картографиране осигуряването и контролът на качеството на дейностите и крайните продукти са въпроси от особена важност. Например, според документацията на проекта IMAGE & CORINE Land Cover 2000, задача 4, "Осигуряване и контрол на качеството", цели да осигури покриването на европейските изисквания от отделните страни, като всяка стъпка в обработката на данните е последвана от съответни проверки [3]. В рамките на услугата за мониторинг на земната повърхност на програмата "Коперник" две от основните дейности са посветени на проблема за тематичната точност - верификация и независима статистичаска валидация.

Настоящият доклад е посветен на проблема за контрол на тематичната точност - елемент на качеството на данните за земното покритие, получени чрез автоматизирана класификация на спътникови изображения на България по програмата "Коперник". Става въпрос за т. н. слоеве с висока разделителна способност (High Resolution Layers, HRL). След като междинните продукти [4], в хода на проекта са били подложени на верификация и подобряване, крайните продукти се валидират за тяхното окончателно приемане. Към датата на изнасяне на този доклад бяха обявени предварителни непълни резултати на интернет страницата на Програмата по биогеографски райони. В нашия случай, на валидация е подложен слоят за степента на непроницаемост (Degree of Imperviousness, съкратено - IMD)) за България от 2012 г. Тематичната точност се определя статистически чрез метода на стратифицирана случайна извадка и сравняване с референтни данни с по-висока пространствена разделителна способност. По отношение на тематичната точност на многоградационните "плътностни" слоеве, каквито са HRL, изследванията, по принцип, са недостатъчни. За България HRL досега са верифицирани главно по метода за визуално-качествено оценяване (look-and feel) [5], а количествен метод за оценка но тематичната точност е приложен само за слоя на водните тела [6]. В настоящата работа по-строго са описани и приложени параметрите на планиране на статистическия подход.

Данни и методи

Използвани данни

Данните, използвани в хода на работата, са геопространствени и могат общо да се разделят на две групи – входни данни и референтни данни.

Входните данни се състоят от самия слой с висока разделителна способност "Степен на непроницаемост". Това са данни в растерен формат, с размер на пиксела 20 m. В този си вид слоят е използван в дейностите по верификация и подобряване от отделните страни-участнички, На валидация, съгласно изискванията на програмата „Коперник”, се подлага интегрираният европейски продукт, който е с размер на пиксела 100 m. Обект на настоящата работа е извършване на независима валидация на продукта [7] за територията на България по собствена методика.

Референтните данни са тестовата основа, с която се сравнява входният слой. Те, като минимум, трябва да отговарят на следните изисквания: да са с по-висока пространствена разделителна способност от оценяваните данни, да им съответстват тематично и да се отнасят към същия времеви период. Важна част от референтните данни са т. н., *in-situ* данни – данни с некосмически, местен произход. Такива са изображенията от цифровата ортофотокарта (ЦОФК) на България, заснети през 2011-12 г. (Таблица 1). Други използвани референтни данни са тези от спътника SPOT 5 във вид на изображения с три спектрални канала.

Таблица 1. Референтни данни

Име/Спътник	ЦОФК	SPOT-5
Брой спектрални канали	3	3
Спектрални канали	R, G, B	G, R, NIR
Пространствена разделителна способност, m	0.4	2.5

Съображения относно подхода за валидация

Целта на валидацията е да бъде получена независима и представителна оценка за тематичната точност на тествания HRL. Проверяваният продукт повече не се променя - при валидацията, за разлика от верификацията, липсва обратна връзка за евентуалното му подобряване. Получената оценка за точността се отразява в метаданните на HRL за информация при по-нататъшното му използване.

Трябва да отбележим, че HRL притежава двойствен (дори многостранен) характер поради факта, че от един многоградационен слой, чрез прагови и други операции, потребителят може да генерира производни слоеве, които съответстват на неговите конкретни потребности. Например, от слоя за плътността на дървесното покритие се получава слоят за типа гора, отговарящ на дефиницията на FAO [4]. В нашия случай се генерира и използва бинарен слой за териториите с плътно застрояване (bult-up areas) чрез сравняване на плътността на непроницаемост по праг 30% (наричан още слой-карта (map layer)). Това е слоят, спрямо който се извършва оценка на тематичната точност [8].

При планирането на работата по валидацията неминуемо възникват въпроси относно характеристиките (качествени показатели) на продукта, например: доколко той съответства на реалността или пък, доколко отговаря на спецификациите. Имайки пред вид двойствения характер на HRL, първият въпрос можем да зададем към слоя с плътностите, а втория – към бинарния слой-карта.

Проверяваният тематичен слой земно покритие за степента на непроницаемост, разглеждан като популация от пиксели, е с голям обем, което налага прилагането на метода на случайната извадка за оценка на тематичната точност. Цялостната оценка на такъв многоградационен продукт се постига чрез тестове в две направления – на бинарен слой за плътно застроени територии (built-up, приемащ стойности от 30 до 100% след прагова операция) и на самия плътностен слой (приемащ целочислени стойности в диапазона 0 – 100, в проценти), базирайки се на идеите, изложени в [8].

Оценка на многоградационния слой

За валидирането на плътностния слой е необходимо да се сравнят неговите стойности за непроницаемост с референтни такива, получени чрез независима оценка. Един вариант за обективна оценка на плътности е чрез налагане на равномерна решетка върху оценявания пиксел [8]. Това е схема с две нива за получаване на случайна извадка. Първо, чрез задаване на местоположението на тестовия пиксел (квадрат с размер 100 m) и второ – чрез проби по равномерната мрежа в него, състояща се от 10x10 точки (фиг. 1). Точките, попадащи върху непропусклива повърхност се маркират с 1, а броят им дава референтната стойност за пиксела в проценти.

Фиг. 1. Измерване на плътността на запечатването чрез равномерна мрежа от 10x10 точки

В резултат на така получените и обработени референтни данни, многоградационният слой за плътността на непроницаемост може да бъде оценен по следните няколко информационни показателя:

- Диаграма на разсейването;
- Параметри на регресионната права;
- Корелационен коефициент.

Диаграмата на разсейването дава обща представа за съгласуваността между продукта и референтните плътности. Параметрите на регресионната линия, съответстваща на множеството отчети предоставя информация за евентуално калибриране на оценявания слой. Корелационният коефициент показва средното отклонение на стойностите на отчетите от тенденцията на разпространение на линията.

Оценка на бинарния слой

Важен въпрос за оценка на качеството е този за коректно задаване на изискванията към продуктите, които се създават в рамките на даден проект. Например, изискването към тематичната точност на слоя за почвеното запечатване от 2006 е било доста приблизително зададено: "Точността на класификацията на хектар (на базата на мрежа 100x100 m) за плътно-застроените и незастроените площи да е най-малко 85% за европейския продукт" [9]. Дали точността ще бъде постигната (и продуктът одобрен), силно зависи от критериите за оценка, които в случая не са прецизирани.

Интервалната оценка на даден параметър на една популация е по-точна от точковата и затова е предпочитан инструмент за статистически анализ [10]. Интервалната оценка за дела на грешката в HRL по случайна извадка от n на брой точки, от които p' са с грешка, е представена чрез формула 1. Членът с квадратния корен от двете страни на неравенствата представлява извадковата грешка. От нея може да се определи изискването за обема на минималната извадка.

$$(1) \qquad p' - z_{\alpha/2}\sqrt{\frac{p'q'}{n}} < p < p' + z_{\alpha/2}\sqrt{\frac{p'q'}{n}}\ ,$$

където: p – действителна грешка (в популацията), n – обем на извадката (брой пробни точки), p' – дял на точките с грешка от неправилна класификация, q' = (1 - p') – дял на благоприятните случаи, z – тестова стандартна стойност, α – ниво на значимост.

Проверка на хипотези

Един по-строг подход към проблема за приемане или отхвърляне на продукта може да се определи от следните две съображения. Първо, решението за приемане на продукта би било оправдано, ако сме сигурни, че действителната тематична точност е над 85% (т.е., действителната грешка е под 15%). И обратно, за да отхвърлим продукта трябва да сме сигурни, че действителната грешка е по-голяма от 15%.

Гореспоменатите съображения могат да бъдат изразени чрез термините на проверката на хипотези (формули 2 и 3). Нулевата хипотеза изразява допускането, че наличната действителна грешка в слоя-карта е под максимално допустимия праг от 15% и тогава продуктът се приема. Обратно, ако действителната грешка е по-голяма или равна на тази максимален праг, продуктът се отхвърля.

(2) H_0: $p < 0.15$;

(3) H_1: $p \geq 0.15$

При работата с хипотези трябва да се отчитат два типа възможни грешки. Ако вярна хипотеза H_0 бива отхвърлена, налице е грешка от тип 1, наричана още риск на производителя. При приемане на невярна хипотеза H_0 се извършва грешка от тип 2, което носи риск на потребителя.

Вероятност за приемане на продукта

Реализирането в популация от N пискела на случайна извадка с обем n, в която k точки са грешни, се описва строго чрез хипергеометрично вероятностно разпределение [10]. При условие, че N е много по-голямо от n (N>10n), то може да бъде апроксимирано с биномно. Ако действителната грешка в слоя обозначим с p, то вероятността в извадката от n пиксела да имаме k грешни има вида (формула 4). Вероятността действителната грешка в слоя да е по-малка от p, се представя чрез формула 5.

(4) $$P_k \approx C_n^k . p^k . (1-p)^{n-k}$$

(5) $$L(p,n,c) = \sum_{k=0}^{c} C_n^k p^k (1-p)^{n-k} ,$$

Величината L, всъщност, е вероятността за приемане на слоя, ако действителната грешка в него е по-малка от p.

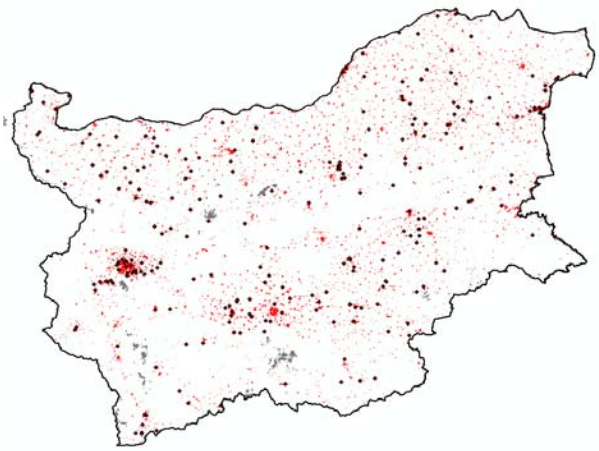
План и получаване на статистическата извадка

При определяне на точността на единствен клас (в случая – HRL за степен на непроницаемост) най-информативни са оценките за двата вида грешки – неправилна класификация (commission) и пропуск (omission). Предварителната оценка за относителната площ на плътното застрояване по входните данни спрямо територията на страната е 2.63%, което налага прилагането на стратификация при оценяването на грешките от тип пропуск. Оценяват се и двата споменати вида грешки и за целта са необходими най-малко две страти, по една за всеки вид грешка:

- Страта 1, страта за грешки неправилна класификация - включва териториите със стойности на непрпускливостта 30% – 100%, т.е., площите с плътно застрояване, според класификатора;
- Страта 2, страта за грешките от тип пропуск - обхваща териториите със стойности 1% - 29% във въпросния HRL.

На базата на предварителна информация за начина на подобряване, приложен преди към този HRL, считаме, че вероятността за намиране на почвено запечатване вън от тези две страти е пренебрежимо малка.

Извадковата грешка, доверителният интервал и обемът на извадката са свързани помежду си (формула 1). От гледна точка на компромис между ресурса време и надеждността на оценката, е избран обем на извадките от по 250 случайни точки за всяка страта. Вариантът за Страта 1 е показан на фиг. 2. Точките са сравнително равномерно разпределени върху плътно застроените територии. При доверителен интервал 90% за Страта 1, максималната извадкова грешка е 5.2%, а при доверителен интервал 68.26%, тя е 3.15%. При пресмятане на грешките от тип пропуск се взема пред вид отношението на площите на двете страти [6].



Фиг. 2. Разположение на 250 случайни точки
по територията на страната

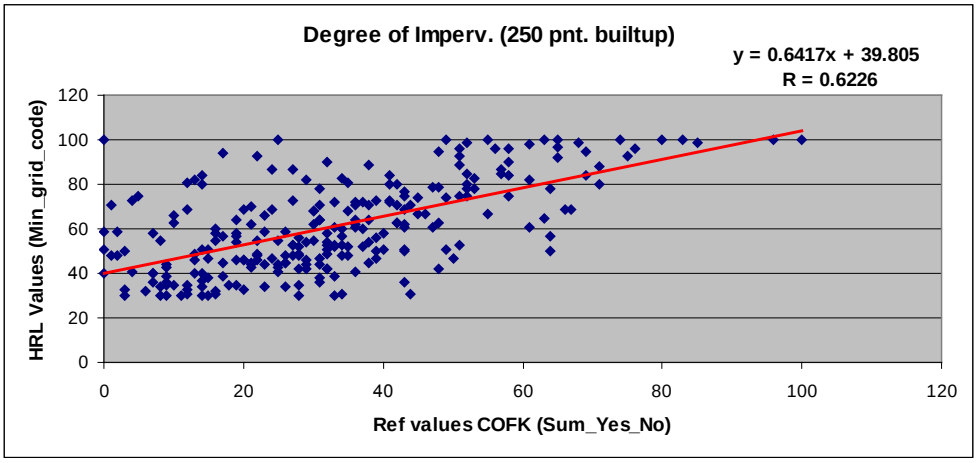
Резултати и обсъждане

В резултат на проведените измервания бяха проверени 250 случайни точки в Страта 1 - тази за плътно застрояване с плътности 30-100%. Стойности на плътността под 30% показаха 116 от тях, което показва, че делът в извадката на грешките от тип "неправилна класификация" е 46.4%.

На фиг. 3 е показана двумерната хистограма от сравняването на референтните и входните плътности на степен на непроницаемост. Коефициентът на корелация е нисък, което показва, че във входните данни са налице значителни отклонения спряма референтните стойности. Наклонът на регресионната права говори за тенденция на завишаване на реалната

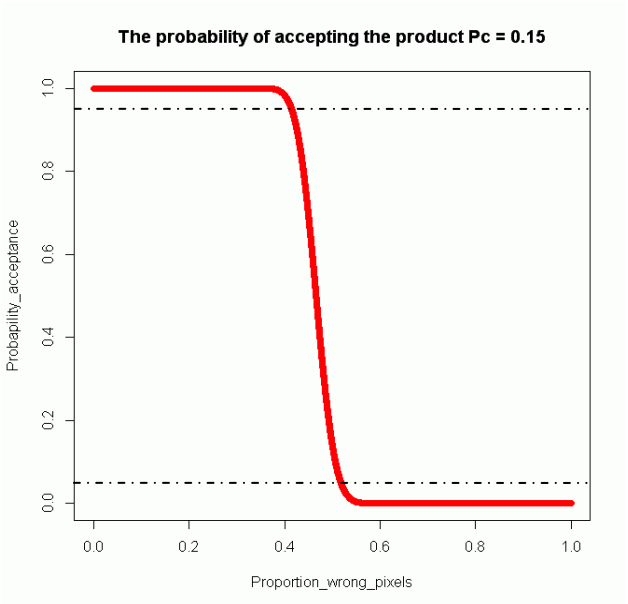
плътност от страна на класификатора. Параметрите на уравнението на линията потенциално могат да се използват за корекция на плътностите с цел калибриране на многоградационния слой. Като цяло, е налице значително разсъгласуване между данните от въпросния HRL и референтните данни. Възможните причини са няколко:

- Геометрично отместване между входните изображение IMAGE2012 и ЦОФК;
- Недостатъчно добра работа на класификатора
- Грешки, причинени от схемата за визуалното дешифриране.



Фиг. 3. Резултати от оценката на многоградационния HRL „Степен на непроницаемост”

На фигура 4 е показана графиката, съгласно формула 5, на вероятността за одобряване на Страта 1 на крайния продукт, в случая – слоя степен на непроницаемост. Доверителният интервал е 90%, което в съгласие с традиционното разбиране за достатъчна надеждност. Границите му са показани с хоризонтални пунктирани линии. Вижда се, че наклонения участък от графиката е вдясно от стойността 0.4 върху абсцисната ос, което говори, че оценката за действителната грешка в слоя е от този порядък. По-конкретно, резултатите са следните: с вероятност 95% може да се твърди, че грешката от неправилна класификация в Страта 1 е над 41.45%. Със същата вероятност се твърди, че същата грешка е под 51.8%. Вероятността, грешката от неправилна класификация в Страта 1 да надхвърли прага от 15%, е 100%.



Фиг. 4. Вероятност за приемане на продукта

Заключение

Цялостните резултати от проведеното изследване демонстрират, възможностите, както на използваните методи за извличане на информация от разновременни спътникови изображения, така и на някои инструменти за оценка на нейното качество. Тестван е многоградационен слой в две посоки – самия плътностен слой (приемащ целочислени стойности в диапазона 0 – 100% и бинарен слой за плътно застроени територии, след прагова операция с праг 30%. Приложените методи за оценка и в двата случая установиха значителни несъответствия между продукта и референтните данни. Продуктът не покрива изискваната тематична точност от 85%. Необходимо е изследванията да продължат в посока на усъвършенстване на методите и алгоритмите за класификация на спътникови изображения за големи територии. Нужно е също и подобряване на методите за проверка на качествените показатели, респективно, на тематичната точност, както на бинарните, така и на многоградационните продукти.

References:

1. Тужаров, Хр. Управление на данни, http://tuj.asenevtsi.com/, 2013.
2. Devillers, R., R. Jeansoulin, eds., Fundamentals of Spatial Data Quality, ISTE Ltd., 2006.
3. CORINE Land Cover update, I&CLC2000 project - Technical Guidelines, EEA, 2002.
4. Langanke, T., GIO land High Resolution Layers (HRLs) – summary of product specifications, EEA, 2013, http://land.copernicus.eu/user-corner/technical-library/gio-land-high-resolution-layers-hrls-2013-summary-of-product-specifications/at_download/file
5. Dimitrov, V. Evaluation of Forest High Resolution Layers 2012 for Bulgaria. CD Proceedings of XXIII International Symposium on Modern Technologies, Education and Professional Practice in Geodesy and Related Fields, Sofia, Bulgaria. No 34, 2013, Publisher: Union of Geodesy and Surveyors.
6. Dimitrov, V., Verification of forest and water high resolution layers 2012 for Bulgaria. CD Proceedings of Ninth scientific conference with international participation “Space, Ecology, Safety” (SES 2013), ISSN 1313-3888.
7. Imperviousness 2012, http://land.copernicus.eu/pan-european/high-resolution-layers/imperviousness/imperviousness-2012/view, последно посетен на 30.01.2017 г. (изтеглен от интернет страницата на програмата "Коперник"през м. юни 2016 г.)
8. Maucha, G., G. Büttner, B. Kosztra. European Validation of GMES FTS Soil Sealing Enhancement Data, In Remote Sensing and Geoinformation not only for Scientific Cooperation, pp. 223-238, EARSeL, 2011.
9. Guidelines for verification of high resolution soil sealing layer - Qualitative assessment - Prepared by: C. Steenmans and A. Sousa, EEA, 2007
10. Христов, В., Основи на теорията на вероятностите и математическата статистика с приложение в техниката и икономиката, Техника, София, 1964.
11. Guidelines for verification of high-resolution layers produced under GMES / Copernicus Initial Operations (GIO) land monitoring 2011 – 2013 - Prepared by: G. Büttner, EEA, 2013.