Computers & Geosciences 35 (2009) 1940–1949

Contents lists available at ScienceDirect

Computers & Geosciences

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

Development of a pit filling algorithm for LiDAR canopy height models

Joshua R. Ben-Arie [a], Geoffrey J. Hay [a,*], Ryan P. Powers [a], Guillermo Castilla [a], Benoît St-Onge [b]

[a] Department of Geography, University of Calgary, 2500 University Dr. NW, Calgary, AB, Canada T2N 1N4
[b] Département de Géographie, Université du Québec à Montréal, Case postale 8888, succursale Centre-ville, Montréal (Québec), Canada H3C 3P8

A R T I C L E I N F O

Article history:
Received 7 June 2008
Received in revised form
25 January 2009
Accepted 20 February 2009

Keywords:
LiDAR
Data pits
Noise removal
Canopy height model
Interactive data language (IDL)

A B S T R A C T

LiDAR *canopy height models* (CHMs) can exhibit unnatural looking *holes* or *pits*, i.e., pixels with a much lower digital number than their immediate neighbors. These artifacts may be caused by a combination of factors, from data acquisition to post-processing, that not only result in a noisy appearance to the CHM but may also limit semi-automated tree-crown delineation and lead to errors in biomass estimates. We present a highly effective semi-automated pit filling algorithm that interactively detects data pits based on a simple user-defined threshold, and then fills them with a value derived from their neighborhood. We briefly describe this algorithm and its graphical user interface, and show its result in a LiDAR CHM populated with data pits. This method can be rapidly applied to any CHM with minimal user interaction. Visualization confirms that our method effectively and quickly removes data pits.

1. Introduction

Light detection and ranging (LiDAR) is an active remote sensing technology that emits pulses of near infra-red light and records the backscatter, resulting in a three-dimensional (3D) point cloud. Once collected, this raw point cloud is typically filtered and classified into first and last returns. When captured over a forest environment, the first returns correspond to the energy echoed from the uppermost vegetation layer of a canopy. Once classified, these points are interpolated to a *digital surface model* (DSM) representing surface elevation above sea level. The last returns correspond to the last detectable signal when a pulse is intercepted by an opaque object, normally the ground (St-Onge et al., 2003). These classified points are interpolated into a *digital terrain model* (DTM), which represents bare terrain elevation above sea level. When the DTM is subtracted from the DSM, the result is a *canopy height model* (CHM), which represents absolute canopy height above the terrain surface.[1] As a result, LiDAR technology allows for large-area acquisition of unprecedented 3D structural forest information. However, such data are not without their challenges.

Data *pits* are typically visible in raster CHMs as (apparently) randomly distributed dark holes that are digitally represented by exceptionally lower digital height values than their neighbors. It is believed that these artifacts are caused by a combination of factors, from data acquisition to post-processing, though no specific cause has been defined in the literature.

It is important to distinguish between data pits and canopy gaps. Gaps are natural openings in a forest canopy that range in size (Spies and Franklin, 1989). However, canopy gaps are not simply bare earth; rather they are often populated with shrubs and saplings at different stages of growth. In fact, canopy gaps are expected when studying forests and are integral to a healthy and dynamic ecosystem (Whitmore, 1989). It is easy to visually discriminate small canopy gaps from data pits. Canopy gaps are asymmetrical, are generally composed of many pixels (even small canopy gaps), and appear 'natural' in a forest landscape. Pits exhibit none of these characteristics. In Fig. 1, the large dark mass (image lower right) is a natural canopy gap, while the scattered small dark rectangles (more than one pixel) and squares (single pixels) represent data pits.

In addition to lending a poor visual appearance to an image (Fig. 1), analyzing a pit-filled dataset may produce inaccurate biophysical and ecological measurements (elaborated below). The challenge is that not all pits are the same value, or differ the same amount in value relative to their neighbors, thus using a simple threshold to define pits does not work. If a group of contiguous pixels corresponding to a tree crown contains a pit consisting of a single pixel, that pixel value can still be higher than many non-pit pixels in other areas of the CHM.

If one were to simply apply smoothing filters – common in remote sensing image processing software – to a CHM with data pits, pits could certainly be removed, or at least their 'influence' in the image visually reduced. However, based on current methods,

* Corresponding author.
E-mail address: gjhay@ucalgary.ca (G.J. Hay).

[1] The common term *digital elevation model* (DEM), as used later in this paper, represents any type of digital model composed of elevation or height values, thus DTMs, DSMs and CHMs are specific types of DEMs.

0098-3004/$ - see front matter Crown Copyright © 2009 Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.cageo.2009.02.003

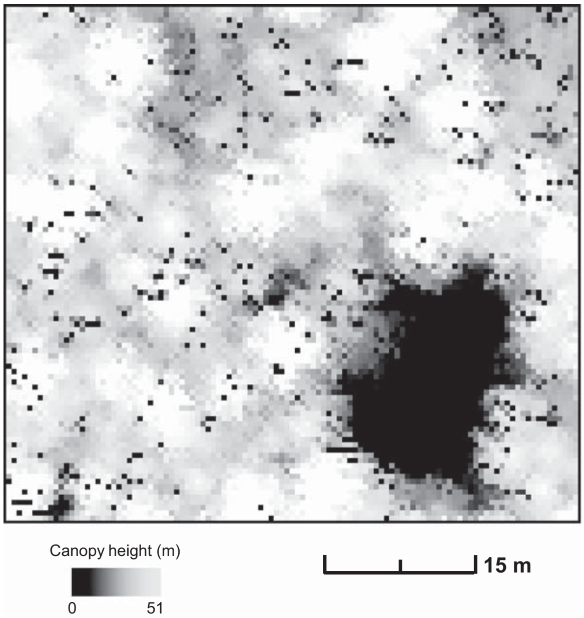

Fig. 1. Subset of Campbell River canopy height model (CHM). Comparing a canopy gap approximately 10 m wide (lower right) to data pits (small dark rectangles and squares).

all pixels in the dataset would also be altered, not just pits. Not only would this result in a visual smoothing of the image – dependent on the size and type of smoothing filter selected – but it is also an inefficient use of LiDAR technology, as smoothing would also alter the vertical-accuracy (i.e., height) of the image. As Hyyppä et al. (2000) reports, forest LiDAR data has an average vertical standard error of approximately 22 cm (varying with the technology used). Thus, when considering LiDAR's numerous and growing applications in forestry and landscape ecology (Dubayah and Drake, 2000; Hyde et al., 2005; Lefsky et al., 2002; Lim et al., 2003; Reutebuch et al., 2005; Hilker et al., 2008; St-Onge et al., 2008), it is imperative to obtain accurate CHMs, especially for individual tree-crown delineation (Leckie et al., 2003) and tree height estimation (Popescu and Wynne, 2004). Additionally, the presence of pits will result in average canopy height under-estimation, which could lead to measurement errors of above-ground forest biomass and carbon sequestration.

The *purpose of this study* is to report on a pit filling method for LiDAR CHM's and to evaluate its results by visually comparing them to the effect common smoothing filters have on the same datasets. As Wood and Fisher (1993) discuss, visual assessment of CHMs is important. Indeed, it is an effective and often under-estimated method in assessing data quality. In the following sections, we provide a more thorough review of data pits within the literature and discuss their potential causes; describe the study site and data; provide details on the methods developed; discuss the results; and provide a conclusion to our work.

2. Background

In this section, we briefly review pertinent literature related to LiDAR pits and discuss their possible causes.

2.1. Insight from the literature

There is a distinct paucity in the literature regarding data pits in LiDAR digital elevation models (DEMs). Leckie et al. (2003) states that these artifacts arise from "ground hits within a tree crown". Their pre-processing attempt to remove pits involved overlaying a 25 cm grid on a 3D point cloud and building a DSM from the highest hit in each cell. However, some pits remained, causing "artifacts in the surface model". Hyyppä et al. (2000) describe some pixels as "no data", and filled them "by using interpolation and a knowledge of near-by pixels." The interpolation method is not given, neither is a definition of these problematic pixels. Kraus and Pfeifer (1998), in their influential paper on DTM production, encountered "big negative blunders", and it is relatively clear that data pits are their equivalent. These "blunders" are not described in any detail, but they did 'tweak' their technique to be more robust in the presence of this problem.

Zhang and Chen (2003), in attempting to remove non-ground LiDAR data points for DTM production, mention that some points have "large negative elevation values drastically lower than those of their neighbors". They also call this phenomenon "negative blunders". To fix this problem, a morphological filter was suggested, but not implemented or tested. They also mentioned that the source of negative blunders is not definitively known.

MacMillan et al. (2003) attempted to fill pits in a LiDAR DTM by applying mean filters of varying sizes. They considered this a sub-optimal approach and called for the development of a more flexible DEM editor. There has been other research conducted on closing pits or minimizing errors in LiDAR DTMs (Briese et al., 2002; Younan et al., 2002). However, these papers do not elaborate on the relative number of pits in the study area or the amount that pits drop in value relative to their neighbors, nor do they explicitly define data pits or their causes. Additionally, there is scarce mention in the literature specifically on pits occurring in LiDAR DEMs, let alone the appropriate filling of these pits. In fact, not a single paper was found devoted to the subject of data pits in LiDAR DEMs. The issue of data pits in LiDAR CHMs is far from resolved.

2.2. Causes of data pits

While not specifically defined, it is probable that the cause(s) of data pits are intricately linked to LiDAR technology and the pre-processing of raw point-clouds into meaningful raster DEMs. Thus, the *raison d'être* of data pits is complex and not fully resolved. The process from laser scanning to raster model production is a multi-step procedure uniting three different technologies: (i) laser ranging system, (ii) inertial navigation system (INS) and (iii) differential global positioning system (DGPS). Because of the possible compounding effect of multiple factors, it is difficult to measure the effect individual factors may have on pit formation.

Range error, platform attitude variation, position accuracy and time misregistration can all contribute to errors in a LiDAR dataset. Attitude errors arise from increasing the flying height and scan angle (Baltsavias, 1999). Time misregistration between the laser instrument, INS and DGPS account for inaccurate 3D positioning (Baltsavias, 1999; Latypov, 2005). Position accuracy, mainly concerning the quality of DGPS processing, accounts for most of the intrinsic technological error associated with LiDAR use (Baltsavias, 1999; St-Onge et al., 2003). However, it is not clear whether system errors produce data pits.

Leckie et al. (2003) postulate that data pits may form by combining different flight line datasets. LiDAR spatial resolution may be affected or unevenly distributed by variation in aircraft attitude, scan pattern, structure of the canopy and terrain and deflection of lost returns. Some areas can yield zero to more than ten pulse returns per square meter, resulting in a failure to record many treetops (Gaveau and Hill, 2003, St-Onge et al, 2003).