Open Data Cube Products Using High-Dimensional Statistics of Time Series

Dale Roberts*, Bex Dunn†, Norman Mueller†
*Australian National University †Geoscience Australia

Abstract—**We describe some new techniques based on high-dimensional statistics of time series to create continental-scale products from Earth Observation data. These methods are developed on the Australian Open Data Cube called 'Digital Earth Australia'. We provide an overview of these new 'second order' statistical algorithms and show how they can be used in various ways.**

I. Introduction

Open Data Cubes[1] provide an architecture to store and analyse deep time series of spatial data. As of early 2018 the Australian implementation of Open Data Cube, Digital Earth Australia (DEA) (previously known as the Australian Geoscience Data Cube) [1], contains the Australian archive of Landsat data from 1986 to 2017, and Sentinel-2 data from 2015 to 2017, along with complementary spatial datasets such as digital elevation and climate data.

To effectively exploit the large volumes of data contained within DEA, we have been developing new techniques to summarise and analyse this data using the philosophy that the high-dimensional nature of the time series data should not be broken by applying one-dimensional methods in each band separately. Our first implementation of this idea was carried out in [2] where a new algorithm to construct cloud-free Pixel Composite Mosaics (PCMs) that maintain all the relationships between the spectral bands was proposed.

Pixel composites generated from Earth Observation data are used for several purposes. For example they are used to provide minimum-cloud images over regions requiring more than a single scene, to create seasonal or annual images of regions to represent specific environmental conditions such as when vegetation is greenest, underpin land cover classification, or provide a basis for change detection [3]–[6].

This paper provides an introduction to some of the new 'second order' statistical techniques that follow and show how we are using these methods for environmental characterisation, change detection, and large-scale calibration of data reduction methods such as the Tasselled Cap transformation [7].

II. Methods

All the algorithms that we present follow the same core methodology: the analysis is done on a (multidimensional) time series basis. More precisely, we consider our space-time stack of observations over the continent as a collection of time series, where $\mathbb{X}_{ij}$ is the time series of observations at pixel location (i,j). At each pixel location, we denote by $N := N_{ij}$ the number of observations (that varies from pixel location to pixel location). The $(i,j)^{\text{th}}$ time series is given by

$$\mathbb{X}_{ij} = [\mathbf{x}_{ij}^{(1)}, \mathbf{x}_{ij}^{(2)}, \cdots, \mathbf{x}_{ij}^{(N)}]^T,$$

and for clarity of exposition, we now omit the "ij" subscript and write $\mathbf{x}^{(t)}$ for the pixel observation at time t and $\mathbb{X}$ the time series at location (i,j).

A. Geometric Median PCM

We briefly recall the geometric median PCM from [2]. For each pixel time series $\mathbb{X}$, we perform the following. Given the p-band pixel observations through time $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)} \in \mathbb{R}^p$, we take the *geometric median* (GM) of these observations defined as

$$m := \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^p} \sum_{t=1}^{N} \|\mathbf{x} - \mathbf{x}^{(t)}\|, \tag{1}$$

where $\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^p$. We place this p-dimensional value m at location (i,j) in our PCM and continue to the next pixel. Our method also involves some ancillary steps and modifications to handle missing values and to mask observations based on pixel quality (cloud, shadow, saturation, etc.) to sneak below the "breakdown point" of the GM, see [2]. The approach works due to the GM's robustness to outliers and skewness.

B. Median Absolute Deviation PCMs

The median absolute deviation (MAD) is a well-known robust measure of the variability of a univariate sample of data that was first proposed by Gauss in 1816 [8]. For a univariate data set $X = \{x_1, x_2, ..., x_n\}$, the MAD is defined as the median of the absolute deviations from the data's median: $\text{MAD} = \text{median}\left(\{|x_i - \text{median}(X)|, i = 1, \ldots, n\}\right)$. To continue the philosophy that the high-dimensional nature of the data should not be broken, we consider two multidimensional variations on this idea. For each pixel time series $\mathbb{X}$ we have the p-band pixel observations through time $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)} \in \mathbb{R}^p$. We then calculate the GM m and consider the *Euclidean MAD* given by

$$\text{EMAD} = \text{median}\left(\left\{\left\|\mathbf{x}^{(t)} - m\right\|_{\mathbb{R}^p}, t = 1, \ldots, N\right\}\right)$$

and the *Spectral MAD* given by

$$\text{SMAD} = \text{median}\left(\left\{\text{cosdist}(\mathbf{x}^{(t)}, m), t = 1, \ldots, N\right\}\right)$$

[1] https://www.opendatacube.org/

where $\text{cosdist}(u,v) := 1 - u \cdot v / (\|u\| \|v\|)$ for $u, v \in \mathbb{R}^p$. We note that since the Euclidean and cosine distances return scalars, 'median' is thus a one-dimensional median and the resulting values of EMAD and SMAD are positive scalars. This means that the applying EMAD and SMAD to every (i,j) pixel time series produces 1-band PCMs.

C. Tasselled Cap Transformation

Tasselled Cap Transformation (TCT) is a data reduction method originally proposed in [7], [9] for Landsat observations that reduces 6 bands (BLUE, GREEN, RED, NIR, SWIR1, SWIR2) to 3 bands that are more interpretable: TC1 (Brightness), TC2 (Greenness), TC3 (Wetness). The TCT is linear and, as such, its "calibration" is given by well-chosen 3×6 matrix Q. The approach has generally been to obtain some cloud-free observations from a given sensor, correct to top-of-atmosphere (or at-satellite) reflectance, perform a principal component analysis (PCA), followed by a Procrustes rotation (PR) to rotate the PCA axes so that they represent interpretable quantities. For example, see the methodology in [10].

Calibrations found in the literature have been conducted on limited spatial regions and number of observations. The resulting calibrations are useful for the area and land cover where the transform was formulated but varies in applicability to other areas and land covers. Effort has been devoted to improving the formulation and calibrating it for other sensors, broader regions, and land covers [10]–[12]. Further, a calibrated TCT should be applied to like-for-like data. That is, a top-of-atmosphere reflectance calibration should not be applied to surface reflectance data. For these reasons, their applicability to our continental-scale surface reflectance data in the DEA was not optimal.

As such, we are proposing a new methodology whereby we calibrate a TCT to our continental PCMs generated from Landsat data corrected to surface reflectance [13]. This new calibration is more representative of the conditions of the entire Australian continent through time and the data in our instance of the Open Data Cube [1]. As this methodology is data independent, it applies to any consistent data appropriate to Open Data Cubes, such as time series of Sentinel-2 L1C data.

III. Products and Discussion

A. Pixel composite mosaics and their application

Using DEA and our methodology, we have generated annual PCMs for each year of available data from 1986 to 2017 from the Landsat-5, -7 and -8 satellites where the data is corrected to surface reflectance using the method of [13] and spatially calibrated to under one pixel accuracy [1]. This work provides a set of high-quality seamless and cloud-free continental mosaics representing annual conditions that serve as baselines for further analyses and are available for free to the public. This makes it possible to perform change detection, functional calibrations, and becomes a useful feature (covariate) for further downstream machine learning applications such as land cover classification on a continental scale. An example subset of a PCM is given in Figure 1.

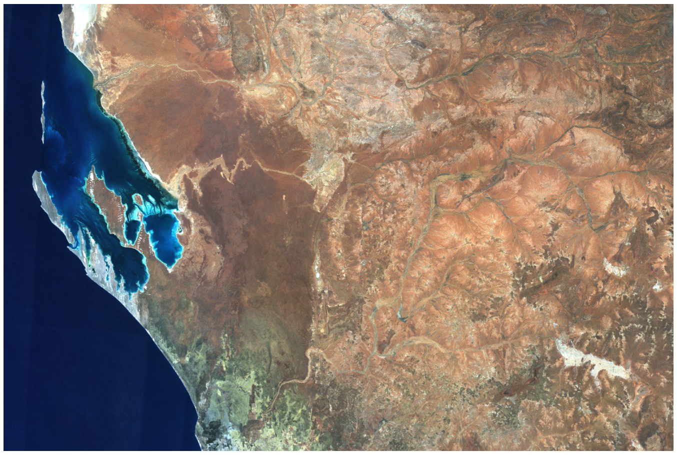

Fig. 1. Subset of continental-scale Geometric Median PCM covering the Gascoyne River catchment in Western Australia, derived from Landsat 8 from 2013 to 2017, displayed in True Color.

B. High-dimensional median absolute deviations, change detection, and land classification

Our main motivation to use high-dimensional statistics to construct PCMs is that it provides a characterisation of a time series of observations that can be compared to a similarly constructed PCM over another time period. Due to the properties of our approach, there are a number of ways this can be achieved. First, since the relationships between bands are maintained in our geometric median PCM, all the classic remote sensing image processing techniques are applicable. For example, given two PCMs over two disjoint temporal epochs (e.g. 2014 and 2015, or Winter and Summer) classic normalised difference band ratios such as the 'normalised difference vegetation index' (NDVI) $\text{NDVI} = (\text{NIR} - \text{Red})/(\text{NIR} + \text{Red})$ can be computed across each PCM, and simple differences between them computed to see the change in NDVI between the epochs. Second, the EMAD and SMAD PCM products can be used in the change detection method to standardise the change. One way is to calculate the cosine distance between the PCMs over two epochs and then divide the resulting PCM by the pooled SMAD. This means that change occurring in areas commonly experiencing change (e.g., cropping areas) are penalised in the resulting PCM and change occurring in areas that are expected to be invariant are highlighted. Finally, the EMAD and SMAD have some interesting properties for land cover classification as the EMAD captures albedo shifts while the SMAD is invariant under albedo shifts. This means that the SMAD is useful for highlighting areas of land cover change within the epoch of the PCM without highlighting areas of significant cloud cover over time. The SMAD also has application for the detection and monitoring of water targets, as water has high variation in reflectance due to physical conditions such as wave action. As such, the SMAD PCM can be combined with a geometric median PCM to highlight these areas, e.g., see Figure 2.

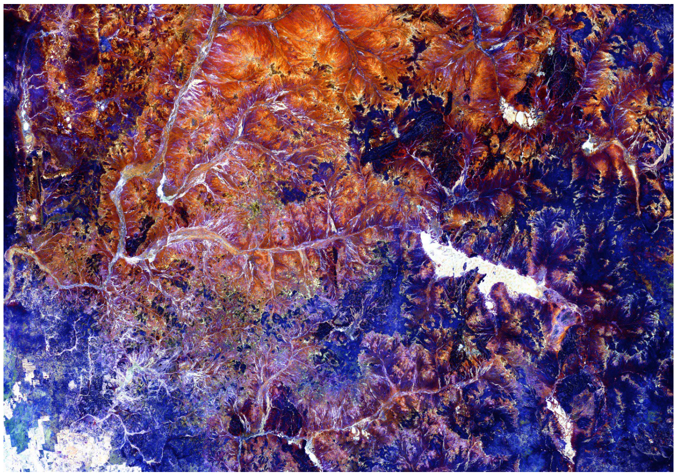

Fig. 2. SMAD enhancement of a subset of Figure 1. The image is coloured by adding the SMAD as a brightness to the continental-scale Geometric Median PCM. Brighter areas have a higher SMAD and hence have shown the greater degree of change over the 2013-2017 epoch.

TABLE I
TCT calibrated to Landsat 8 surface reflectance using our continental-scale methodology.

	BLUE	GREEN	RED	NIR	SWIR1	SWIR2
TC1	0.3567	0.3567	0.3567	0.5350	0.5350	0.2140
TC2	0.1987	-0.2826	-0.2724	0.5357	0.2388	-0.6800
TC3	0.5702	0.1584	0.2627	-0.3959	-0.0045	-0.6511

TABLE II
TCT calibrated to Landsat 8 at-satellite reflectance from [10].

	BLUE	GREEN	RED	NIR	SWIR1	SWIR2
TC1	0.3029	0.2786	0.4733	0.5599	0.5080	0.1872
TC2	-0.2941	-0.2430	-0.5424	0.7276	-0.0713	-0.1608
TC3	0.1511	0.1973	0.3283	0.3407	-0.7117	-0.4559

C. Calibrating the TCT and its application

We calibrated a TCT on the a continental-scale Landsat 8 PCM of surface reflectance data in DEA (shown in Figure 1). The resulting coefficients are given in Table I. We chose to calibrate to a single sensor PCM for comparison to the Landsat 8 coefficients published in [10], given in Table II. Our PCM-calibrated TCT matrix is largely similar in magnitudes but with a number of important differences such as sign flips and slight variations of loadings. We tested this new TCT calibration in a study of the behaviour of Groundwater Dependent Ecosystems (GDEs), focusing on several wetland areas across Northern Australia. By characterising the spatial and temporal changes of GDEs we can establish the dynamics of these ecosystems and quantify the impacts of rainfall and groundwater use. For this study we use the two TCT calibrations given in Table II and Table I. Each TCT was thresholded and then new composite images constructed summarising how many times thresholds for brightness, greeness and wetness were exceeded compared to the number of available observations. Example results for Arafura swamp in Arnhem Land in Australia’s Northern Territory are shown in Figure 3 for the previously published coefficients, and Figure 4 for the newly calibrated coefficients. These show that the newly calibrated TCT coefficients have less overlap between ‘wet’, ‘bright’ and ‘green’ than the previously published coefficients [10]. Further, a number of pixels are identified by the published TCT as being both ‘green’ and ‘wet’ throughout the summary, and there are few areas that are identified as ‘bright’ that are not also identified as ‘green’. This leads to difficulty in picking out areas of change. Conversely in Figure 4 some of the small water bodies on the plateau above the swamp (left side of image) alternate between ‘wet’ and ‘bright’ (visible in magenta) where others are consistently ’wet’ (visible in blue). In comparison in Figure 3 it is harder to distinguish the water bodies that dry seasonally from those that are persistent as they are largely cyan (i.e., both ’green’ and ’wet’). This can be attributed to the published TCT [10] result leading to misidentification of areas of water and bare soil as ’green’. In general the recalibrated coefficients are providing better discrimination between ’bright’, ’green’ and ’wet’ for visualisation and analysis. Consequently our recalibrated coefficients are being used to progress the study.

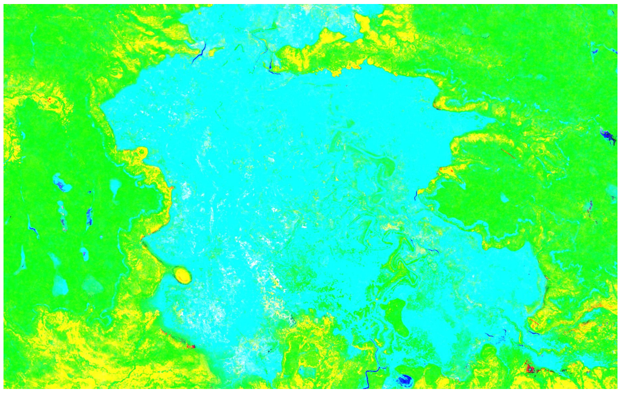

Fig. 3. Summary image of TCT for the Arafura Swamp area during the epoch 2014 to 2017 created using the previously published calibration of TCT given by Table II. The figure is displayed in RGB composite, with Bright shown in red, Green shown in green and Wet shown in blue. Where multiple conditions have occurred over the epoch, colours combine. Eg. cyan colours indicate areas having experienced both Wet and Green over the epoch. Note the lack of contrast between the TCT components, resulting from the older coefficient calibration being less suitable for the study area.

IV. Conclusion

The application of high-dimensional statistics to earth observation data in the Open Data Cube environment enables the time dimension to be used as an extension of the spectral data, rather than just mass analysis of individual observations. This delivers several advantages to environmental analysis, such as effective representation of a time series of observations as a single, characteristic composite, mathematically rigorous change detection, and implementation of classic earth observation transforms on continental time series. The method is not bound to specific data types or processing levels, and has been demonstrated on Landsat and Sentinel-2 data, including those processed to basic orthorectified levels as well as to surface reflectance. We have demonstrated how

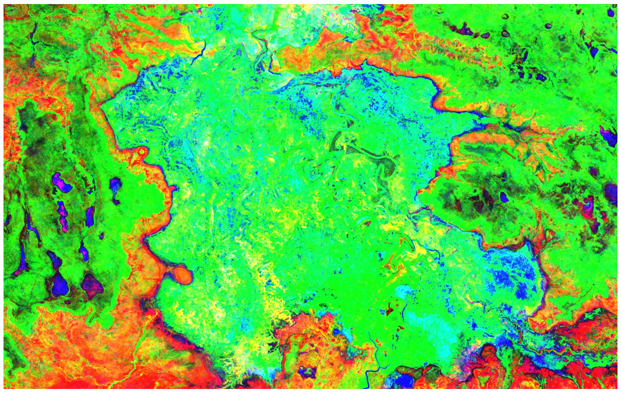

Fig. 4. Summary image for the Arafura Swamp area during the epoch 2014 to 2017 created using the proposed calibration of TCT given by Table I. The figure is displayed in RGB composite, with Bright shown in red, Green shown in green and Wet shown in blue. Where multiple conditions have occurred over the epoch, colours combine. Eg. magenta colours indicate areas having experienced both Wet and Bright over the epoch, indicating ephemeral water bodies. Note the improved contrast between the TCT components compared to Figure 3, resulting from the coefficient recalibration in the proposed method.

these techniques can be applied to image compositing, change detection, Principal Components Analysis and the Tasselled Cap transform, and expect that the further application of high-dimensional statistics in this fashion will enable many more complex analyses including fire scar and water mapping as well as land cover mapping more generally.

V. Acknowledgements

This paper was published with the permission of the CEO, Geoscience Australia.

References

[1] A. Lewis, S. Oliver, L. Lymburner, B. Evans, L. Wyborn, N. Mueller, G. Raevski, J. Hooke, R. Woodcock, J. Sixsmith, W. Wu, P. Tan, F. Li, B. Killough, S. Minchin, D. Roberts, D. Ayers, B. Bala, J. Dwyer, A. Dekker, T. Dhu, A. Hicks, A. Ip, M. Purss, C. Richards, S. Sagar, C. Trenham, P. Wang, and L.-W. Wang, "The australian geoscience data cube - foundations and lessons learned," *Remote Sensing of Environment*, vol. 202, no. Supplement C, pp. 276–292, 2017, big Remotely Sensed Data: tools, applications and experiences. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0034425717301086

[2] D. Roberts, A. McIntyre, and N. Mueller, "High-dimensional pixel composites from earth observation time series," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. issue 11, pp. 6254–6264, 2017.

[3] N. Flood, "Seasonal composite landsat tm/etm+ images using the medoid (a multi-dimensional median)," *Remote Sensing*, vol. 5, no. 12, pp. 6481–6500, 2013.

[4] J. C. White, M. A. Wulder, G. W. Hobart, J. E. Luther, T. Hermosilla, P. Griffiths, N. C. Coops, R. J. Hall, P. Hostert, A. Dyk *et al.*, "Pixel-based image compositing for large-area dense time series applications and science," *Canadian Journal of Remote Sensing*, vol. 40, no. 3, pp. 192–212, 2014.

[5] T. Hermosilla, M. A. Wulder, J. C. White, N. C. Coops, and G. W. Hobart, "An integrated landsat time series protocol for change detection and generation of annual gap-free surface reflectance composites," *Remote Sensing of Environment*, vol. 158, pp. 220–234, 2015.

[6] P. Griffiths, S. van der Linden, T. Kuemmerle, and P. Hostert, "A pixel-based landsat compositing algorithm for large area land cover mapping," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 6, no. 5, pp. 2088–2101, 2013.

[7] R. Kauth and G. Thomas, "The tasselled cap — a graphic description of the spectral-temporal development of agricultural crops as seen by landsat," *LARS Symposia, paper 159*, 1976.

[8] C. F. Gauss, "Bestimmung der genauigkeit der beobachtungen," *Zeitschrift für Astronomie und verwandt Wissenschaften*, vol. 1, pp. 187–197, 1816.

[9] E. P. Crist and R. C. Cicone, "A physically-based transformation of thematic mapper data—the tm tasseled cap," *IEEE Transactions on Geoscience and Remote sensing*, no. 3, pp. 256–263, 1984.

[10] M. H. A. Baig, L. Zhang, T. Shuai, and Q. Tong, "Derivation of a tasselled cap transformation based on landsat 8 at-satellite reflectance," *Remote Sensing Letters*, vol. 5, no. 5, pp. 423–431, 2014.

[11] E. P. Crist, "A tm tasseled cap equivalent transformation for reflectance factor data," *Remote Sensing of Environment*, vol. 17, no. 3, pp. 301–306, 1985.

[12] C. Huang, B. Wylie, L. Yang, C. Homer, and G. Zylstra, "Derivation of a tasselled cap transformation based on landsat 7 at-satellite reflectance," *International Journal of Remote Sensing*, vol. 23, no. 8, pp. 1741–1748, 2002.

[13] F. Li, D. L. Jupp, M. Thankappan, L. Lymburner, N. Mueller, A. Lewis, and A. Held, "A physics-based atmospheric and brdf correction for landsat data over mountainous terrain," *Remote Sensing of Environment*, vol. 124, no. Supplement C, pp. 756–770, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0034425712002544