EVALUATION OF RANDOM FOREST–BASED ANALYSIS FOR THE GYPSUM DISTRIBUTION IN THE ATACAMA DESERT

D. Hoffmeister[1,]*, M. Herbrecht[1], T. Kramm[1], P. Schulte[2]

[1] Institute of Geography, GIS & RS Group, University of Cologne, Albertus-Magnus-Platz, 50923 Cologne, Germany- (dirk.hoffmeister@uni-koeln.de)
[2] Department of Geography, RWTH Aachen University, 52056 Aachen, Germany

KEY WORDS: GIS, remote sensing, random forest, satellite imagery, indices, digital soil mapping

ABSTRACT:

Gypsum-rich material covers the hillslopes above ~ 1000 m of the Atacama and forms the particular landscape. In this contribution, we evaluate random forest-based analysis in order to predict the gypsum distribution in a specific area of ~ 3000 km², located in the hyperarid core of the Atacama. Therefore, three different sets of input variables were chosen. These variables reflect the different factors forming soil properties, according to digital soil mapping. The variables are derived from indices based on imagery of the ASTER and Landsat-8 satellite, geomorphometric parameters based on the Tandem-X World DEM™, as well as selected climate variables and geologic units. These three different models were used to evaluate the Ca-content derived from soil surface samples, reflecting gypsum content. All three different models derived high values of explained variation ($r^2 > 0.886$), the RMSE is ~ 4500 $mg \cdot kg^{-1}$ and the NRMSE is ~ 6%. Overall, this approach shows promising results in order to derive a gypsum content prediction for the whole Atacama. However, further investigation on the independent variables need to be conducted. In this case, the ferric oxides index (representing magnetite content), slope and a temperature gradient are the most important factors for predicting gypsum content.

1. INTRODUCTION

The hyperarid environment of the Atacama between Chañaral and Arica (N Chile) shows geomorphic processes of remarkable slowness, as postulated by geochronological studies on the age of Atacama landforms and surfaces (Dunai et al., 2005). The stability of surfaces in the central desert is documented by smooth slope morphologies, which result, supported by the presence of Biological Soil Crusts (BSC) (Wang et al., 2017), from the accumulation of thick atmospherically derived salt and dust deposits. In particular, the Coastal Cordillera above ~ 1000 m is covered by powdery, gypsum-rich material (called “chuca”), masking hillslopes in the hyperarid core of the Atacama.

Morphodynamic activity mostly occurs due to fog-related atmospheric moisture (i.e., western Coastal Cordillera) and in relation to episodically occurring Andean discharge (i.e., the Precordillera and alluvial fans of the Central Depression), which is generally linked to severe precipitation events potentially causing overland flow or flash floods even in the hyperarid core of the Atacama. Likewise, salt-driven shrink-swell-, slump-, or solifluction-type processes and seismic shaking are assumed to actively contribute to the evolution of the specific Atacama landscape (May et al., 2019, 2020; Ullmann et al., 2019).

Thus, the distribution of gypsum content in the surfaces of the Atacama is an important parameter for landscape characterisation. Therefore, we test in this contribution random forest-based machine learning (RDF) (Breiman, 2001) in order to predict gypsum content by using parameters derived by remote sensing, such as geology, soil surface indices or geomorphometric conditions. This approach follows the “scorpan”-paradigm from digital soil mapping (McBratney et al., 2003), where a soil property (s) is the function of climate (c), organisms (o), relief (r), parent material (p), age (a) and the spatial position (n). Ultimately, we aim at determining the most important factors, delivered by RDF as well, explaining the spatial pattern of gypsum content, which is assumed to be related to specific surface properties.

2. MATERIALS AND METHODS

2.1 Study area and sampling locations

The study area presented in this contribution (Figure 1) is mainly situated in the Coastal Cordillera north of the Río Loa, with a size of 60 by 50 km, partly covering the inactive salt lake Salar Grande and the active Salar de Llamara. The 30 sampling locations on altering hillslopes were chosen based on a pre-analysis of false-colour composites of Landsat 8 and existing geologic maps.

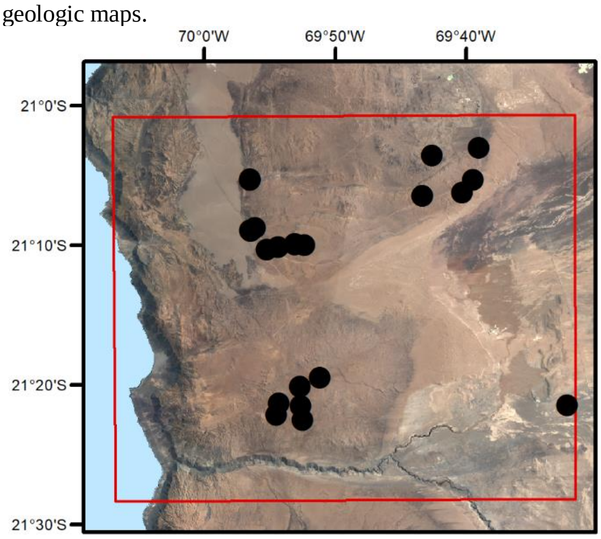

Figure 1. Location of the study area (red rectangle) and the sampling locations (black dots). Background image is a Landsat-8 mosaic derived as an average of images from 2017 by Google Earth Engine.

* Corresponding author

2.2 Control data

Soil surface samples were taken at the presented locations and the inorganic element concentration was determined using an Ametek X-ray Fluorescence Spectrometer (Spectro Xepos, 2007) and 8 g of sample material mixed with 2 g of Cereox analysis wax (Fluxana). Following calculation procedure, root mean square errors of calibration and lower limit of detection can be inferred from Spectro Xepos (2007). Certified reference materials were used for the calibration of the measurements. From these XRF analysis, we used the Ca-content as mg kg^{-1} in the samples for the representation of gypsum content. As expected, the Ca-content (mean: 44977 mg kg^{-1}, SD: 27763 mg·kg^{-1}) shows a high significant correlation with S ($r = 0.98$; $\alpha \leq 0.05$) and a weak contrasting correlation with Na ($r = -0.15$) and Cl ($r = -0.23$), representing sodium chloride.

2.3 Independent variables

The raw digital elevation model (DEM) of the Tandem-X World DEM™ (Kramm and Hoffmeister, 2019) was masked and aggregated to a 60 m resolution. All following indices from satellite images were derived from mosaics of the area, calculated with Google Earth Engine by using the average of all images of a given time frame. For ASTER-based indices, the “ASTER L1T Radiance” dataset were used, consisting of images from the year 2006 that were calibrated to an at-sensor radiance and orthorectified. Likewise, a Landsat 8 mosaic (L8) was computed from the archive “USGS Landsat 8 Collection 1 Tier 1 TOA Reflectance” from 2017, which incorporates a top-of-atmosphere (TOA) calibration.

Independent variables are established by the previously described datasets to represent the different factors of the described “scorpan”-approach (Table 1). Thus, climate (c) is represented by the mean temperature and precipitation derived from the WorldClim 2 dataset (Fick and Hijmans, 2017). BSCs play an important role in covering surfaces in the Atacama and the BSC-Index introduced by Chen et al. (2005) is used here as factor “o”. Relief (r) characteristics are height, slope, landforms from the geomorphons approach and the Topographic Ruggedness Index (TRI). In addition, the distance to sea is calculated. The parent material is represented by several different, sometimes contradicting indices, namely the ASTER-based mineral indices for ferric oxides, kaolinites, SiO_2, gypsum and silicates. Likewise, the surface properties are represented by the ASTER-based Clay Index and the L8-based Grain Size Index (GSI). Most variables are also shown at https://www.indexdatabase.de/ and calculations of each variable are presented in Table 1. Age (a) is not specifically represented here, but is partly regarded by the geologic units. All data is aggregated or resampled to a 60 m resolution as a trade-off between the different sensor resolutions (e.g. DEM resolution is 12.5 m, ASTER TIR-Band resolution is 90 m) and stored in WGS 84 / UTM Zone 19S, EPSG: 32719.

2.4 Statistical methods

A lot more indices were regarded first, but neglected as strong cross-correlation occurred, e.g. by different indices for silica. The random-forest based analysis were conducted in ArcGIS Pro (v. 2.2) using three different sets of inputs for the calculation: 1) all presented datasets as independent variables (“ALL”), 2) a subset of the 10 most important variables from set 1 (“BEST”) and 3) a dataset with only the most important parameter for each of the “scorpan”-factors (“MINIMUM”). The calculation was conducted with 5000 trees, a leaf-size of five and a depth range from nil to nine. The predicted surface was set to a 60 m resolution. Due to the small sample size, no training data was excluded for validation. In contrast, the RMSE and NRMSE (normalized by range) was calculated for the differences between the sample content and the predicted surfaces.

Variable	Formula / Description	Reference
Climate (c)		
Mean temperature		Fick and Hijmans, 2017
Mean precipitation		Fick and Hijmans, 2017
Organisms (o)		
Biological Soil Crust Index (BSCI)	L8: $\frac{1-3* \lvert B4-B3 \rvert}{\frac{B3+B4+B5}{3}}$	Chen et al., 2005
Relief (r)		
Height	From Tandem-X World DEM™	
Slope		
Landform		Jasiewicz and Stepinski, 2013
Topographic Ruggedness Index (TRI)		Riley et al., 1999
Distance to Sea	Orthogonal distance in meters	
Parent material (p)		
Ferric oxides	ASTER: $\frac{\text{SWIR\_B4}}{\text{VNIR\_B3N}}$	Kalinowski and Oliver, 2004
Kaolinites	ASTER: $\frac{\text{SWIR\_B7}}{\text{SWIR\_B5}}$	Hewson et al., 2001
SiO2	ASTER: $\frac{\text{TIR\_B13}}{\text{TIR\_B12}}$	Ninomiya and Fu, 2002
Gypsum	ASTER: $\frac{\text{TIR\_B10} + \text{TIR\_B12}}{\text{TIR\_B11}}$	Cuhady, 2017
Silicates	ASTER: $\frac{\text{TIR\_B11}}{\text{TIR\_B10}}$	Kalinowski and Oliver, 2004
Clay Index	ASTER: $\frac{\text{SWIR\_B5} \times \text{SWIR\_B7}}{\text{SWIR\_B6} \times \text{SWIR\_B6}}$	Kalinowski and Oliver, 2004
Grain Size Index (GSI)	L8: $\frac{B4-B2}{B4+B2+B3}$	Xiao et al., 2006
Age (a)		
Geologic Units	Units of map with scale 1: 1 Mio.	

Table 1. Overview on independent variables with band combination or description and reference, if applicable.

3. RESULTS

Although a small number of samples was used, the results of the three different RDF models (Table 2) show that RDF is suitable to predict Ca-content pretty well. The explained variation for all models is very high ($r^2 > 0.886$) and the RMSE and NRMSE in comparison to the samples is ~ 4500 mg·kg^{-1}, ~ 6% low.

The first model with all parameters (“ALL”) shows already the pattern of important parameters, which is reflected by both other models. The ASTER-based ferric oxide index (representing magnetite content) and the gypsum index, slope and temperature play the most important role (importance > 10%), whereas all other factors play a minor role in the model. Both categorical variables, the geology and the geomorphons based landforms, show the smallest importance and mean annual precipitation has hardly an influence on the results.

The second model (“BEST”) incorporates the ten most important variables of the first model. In this case, a nearly similar order of importance is calculated, with slightly higher overall values. The explained variation is similar to the previous result, but the RMSE and NRMSE is slightly better.

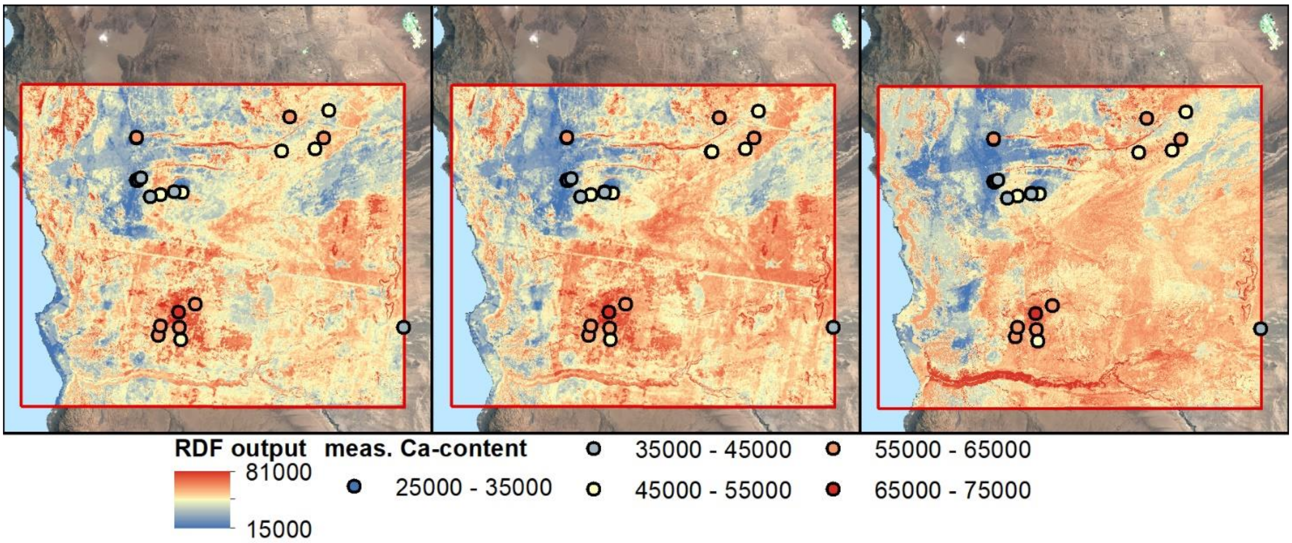

Figure 2. Predicted Ca-content (low: blue colours, high: red colours) for the study area. Left: RDF with all parameters; centre: RDF with the ten best parameters; right: RDF with selected parameters. Sampling locations are coloured equally. Values as mg·kg^{-1}.

The third model ("MINIMUM") regards the most important parameters from the previous models, but only one parameter for each factor of the "scorpan"-approach is chosen. For example, only the ferric oxide index is used to represent the parent material and the gypsum index is neglected. Although, only four parameters are used, the explained variation, RMSE and NRMSE show only slightly higher results.

In Figure 2, all predicted surfaces are presented with a 60 m resolution and the soil sample locations coloured by their Ca-content. The derived distribution fits to the values of the samples and reflects the altering surfaces in the area, e.g. the coastal areas and the Salar Grande (north-western bluish area) with minor Ca-contents. In contrast, hillslopes show higher Ca-contents. Obviously, a small sensor error is distributed in the models, as a linear feature from west to east in the centre section is visible. Differences between the predicted surfaces are for example observable at the coastal slopes in the west and the Río Loa canyon. The first two models derive more areas with higher values (dark red colour) than the last model. In contrast, this last model ("MINIMUM") derives higher values for the Río Loa canyon, as the input of slope is more important in the model.

4. DISCUSSION

Sampling locations were predefined by false-colour images and the geologic map for the area in order to cover different units with varying gypsum content. However, the amount of samples is very low, as sampling is restricted by reachability in this environment. The RDF approach is originally intended for higher sample amounts, but in this case also shows its suitability.

Overall, this simple test of applying RDF worked pretty well and shows a high explained variation ($r^2 > 0.858$), a low RMSE and NRMSE in regard to the small number of samples ($n = 30$) and the large area of ~ 3000 km² covered by these samples.

Variables		Results	
Dataset	Parameters with importance values [%]	Expl. Variation [R²]	RMSE [mg·kg^{-1}]/ NRMSE
ALL	ferric oxides (13), slope (11), temperature (10), gypsum (10), kaolinites (9), SiO_2 (8), height (7), TRI (7), silica (4), GSI (4), clay (4), BSCI (4), distance to sea (4), geomorphons (2), geology (1), precipitation (0)	0.929	4851.24 0.079
BEST	ferric oxides (15), slope (13), gypsum (12), temperature (12), kaolinites (11), SiO_2 (9), height (9), TRI (9), silica (6), GSI (5)	0.924	4260.33 0.042
MINIMUM	slope (27), ferric oxides (27), temperature (26), BSCI (20)	0.886	4591.39 0.065

Table 2. Results of the different RDF models for Ca prediction, computed with different sets of variables, their specific importance (in %), the overall explained variation (as R²), the RMSE (in mg·kg^{-1}) and NRMSE calculated from the sample content and the predicted surface.

The independent variables were chosen from a large number of possible indices, geomorphometric values and further available parameters in order to cover the factors given by the "scorpan"-approach. As described, the derived ASTER image shows an error in the centre of the study area, where no sample points are located. Filtering, atmospheric correction or a more detailed mosaic generation might solve this problem. However, the large size of the Atacama needs cloud-based solutions, instead of locally working with hundreds of single scenes. In addition, the ASTER sensor itself is over the end of lifetime and shows a malfunction of the SWIR sensors since 2007 (e.g. Abrams and Yamaguchi, 2018). Thus, variables should be derived from L8 or Sentinel-2 imagery in the future. Likewise, existing indices based on single band calculations were used as independent variables. Another possibility would be to directly use all band information in order to predict gypsum content.

However, the potential of RDF-based predictions for specific soil surface properties is shown. In particular, the low RMSE is promising. The minimum impact of both categorical variables, landforms and geologic units, might be the result of a bias, as these variables are internally treated by one-hot encoding.

Interestingly, the ferric oxide index (Kalinowski and Oliver, 2004) plays the most important role for all RDF models, as well as slopes and the small temperature gradient, varying from 23.2 °C to 27.8 °C. The ASTER-based gypsum index itself shows a prominent role, but minor results in direct use (r = 0.35). The important role of biological soil crusts in the Atacama, reflected by the BSCI, is not supported here, which might be the result of the index itself. Overall, the results are in accordance with the results of Voigt et al. (2020).

5. CONCLUSION

Random forest-based analysis was tested for the prediction of gypsum content distribution in the Atacama. The first results from three different models show promising results with a high explained variation, a small RMSE and NRMSE. Input parameters for this modelling approach were chosen according to factors from digital soil mapping. Further research needs to exchange ASTER-based indices and extent the small number of samples.

ACKNOWLEDGEMENTS

The project is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 268236062 – SFB 1211. The TanDEM-X WorldDEM™ data is provided by a DLR Science grant, 2017.

REFERENCES

Abrams, M., Yamaguchi, Y., 2018. 20 Years of ASTER Contributions to Lithologic Mapping and Mineral Exploration 1–29.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. doi.org/10.1007/978-3-662-56776-0_10

Chen, J., Yuan, M., Wang, L., Shimazaki, H., Tamura, M., 2005. A new index for mapping lichen-dominated biological soil crusts in desert areas 96, 165–175. doi.org/10.1016/j.rse.2005.02.011

Cuhady, T., 2017. National ASTER Map TIR Gypsum index [WWW Document]. URL https://researchdata.ands.org.au/national-aster-map-gypsum-index

Dunai, T.J., González López, G.A., Juez-Larré, J., 2005. Oligocene-Miocene age of aridity in the Atacama Desert revealed by exposure dating of erosion-sensitive landforms. Geology 33, 321–324. doi.org/10.1130/G21184.1

Fick, S.E., Hijmans, R.J., 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. Int. J. Climatol. 37, 4302–4315. doi.org/10.1002/joc.5086

Hewson, R., Cudahy, T., Huntington, J., 2001. Geologic and alteration mapping at Mt Fitton, South Australia, using ASTER satellite-borne data, in: International Geoscience and Remote Sensing Symposium (IGARSS). pp. 724–726. doi.org/10.1109/IGARSS.2001.976615

Jasiewicz, J., Stepinski, T.F., 2013. Geomorphons-a pattern recognition approach to classification and mapping of landforms. Geomorphology 182, 147–156. doi.org/10.1016/j.geomorph.2012.11.005

Kalinowski, A., Oliver, S., 2004. ASTER Mineral Index Processing Manual.

Kramm, Hoffmeister, 2019. A Relief Dependent Evaluation of Digital Elevation Models on Different Scales for Northern Chile. ISPRS Int. J. Geo-Information 8, 430. doi.org/10.3390/ijgi8100430

May, M.S., Meine, L., Hoffmeister, D., Brill, D., Medialdea, A., Wennrich, V., Gröbner, M., Schulte, P., Steininger, F., Deprez, M., Kock, T. De, Bubenzer, O., 2020. Origin and timing of past hillslope activity in the hyper-arid core of the Atacama Desert – The formation of fine sediment lobes along the Chuculay Fault System , Northern Chile 184. doi.org/10.1016/j.gloplacha.2019.103057

May, S.M., Hoffmeister, D., Wolf, D., Bubenzer, O., 2019. Zebra stripes in the Atacama Desert revisited – Granular fingering as a mechanism for zebra stripe formation? Geomorphology 344, 46–59. doi.org/10.1016/j.geomorph.2019.07.014

McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3–52. doi.org/10.1016/S0016-7061(03)00223-4

Ninomiya, Y., Fu, B., 2002. Quartz Index , Carbonate Index and SiO2 Content Index Defined for ASTER TIR Data. J. Remote Sens. Soc. Japan 22, 50–61.

Riley, S., Degloria, S., Elliot, S.D., 1999. A Terrain Ruggedness Index that Quantifies Topographic Heterogeneity. Int. J. Sci. 5, 23–27.

Ullmann, T., Sauerbrey, J., Hoffmeister, D., May, S.M., Baumhauer, R., Bubenzer, O., 2019. Assessing spatiotemporal variations of sentinel-1 InSAR coherence at different time scales over the atacama desert (Chile) between 2015 and 2018. Remote Sens. 11, 1–22. doi.org/10.3390/rs11242960

Voigt, C., Klipsch, S., Herwartz, D., Chong, G., Staubwasser, M., 2020. The spatial distribution of soluble salts in the surface soil of the Atacama Desert and their relationship to hyperaridity. Glob. Planet. Change 184, 103077. doi.org/10.1016/j.gloplacha.2019.103077

Wang, F., Michalski, G., Luo, H., Caffee, M., 2017. Role of biological soil crusts in affecting soil evolution and salt geochemistry in hyper-arid Atacama Desert, Chile. Geoderma 307, 54–64. doi.org/10.1016/j.geoderma.2017.07.035

Xiao, J., Shen, Y., Tateishi, R., Bayaer, W., 2006. Development of topsoil grain size index for monitoring desertification in arid land using remote sensing. Int. J. Remote Sens. 27, 2411–2422. doi.org/10.1080/01431160600554363