pTITAN2

Permutation of Treatment Labels and Threshold Indicator Taxa ANalysis

Stephanie Figary1, Peter E. DeWitt2,3,4, and Naomi Detenbeck5

1(former) ORISE participant at U.S. Environmental Protection Agency, Atlantic Coastal Environmental Sciences Division, 27 Tarzwell Drive, Narragansett, RI, 02882, USA

2Section of Informatics and Data Science, Department of Pediatrics, University of Colorado School of Medicine, Aurora Colorado, USA

3National Renewable Energy Laboratory, Golden Colorado, USA

4(formerly) Neptune and Company, Inc. Lakewood Colorado, USA

5U.S. Environmental Protection Agency, Atlantic Coastal Environmental Sciences Division, 27 Tarzwell Drive, Narragansett, RI 02882, USA

**Abstract**

Taxa Indicator Threshold ANalysis (TITAN) was developed to identify thresholds along environmental gradients where rapid changes in taxa frequency and relative abundance are observed. TITAN determines separate change-points for increasing and decreasing taxa in aggregate, as well as change-points for individual taxa, with associated confidence intervals generated using bootstrapping. However, if TITAN is applied to different classes of observations, additional analyses besides using non-overlapping confidence intervals are needed to establish whether change-points differ between treatments or groups because non-overlapping confidence intervals can indicate significant differences but overlapping confidence intervals do not necessarily mean the null hypothesis cannot be rejected. To address this, we present a new R package, pTITAN2, which is an extension to the existing TITAN2 package. The pTITAN2 package was developed to enable comparisons of TITAN output between treatments by permutating the observed data between treatments and rerunning TITAN on the permuted data. The output from pTITAN2 can be used to perform the appropriate statistical tests and determine statistical differences without using overlapping confidence intervals.

**Keywords:** TITAN, permutations, thresholds, community composition

# 1 Introduction

Community ecologists are interested in understanding the structure and interactions of multiple species in a given area or habitat type and many are interested in understanding how communities change in response to changing environmental or anthropogenic gradients. One method a community ecologist can use for understanding to detect changes or ecological thresholds across environmental gradients is Taxa Indicator Threshold ANalysis (TITAN) (Matthew E. Baker and King 2010). TITAN is useful for determining the impacts of environmental or anthropogenic gradients such as upstream impervious cover (IC) in a watershed in community ecology studies because it both analyzes each individual taxa response and the community as a whole in the same analysis. Additionally, unlike other community ecology methods, TITAN separates the taxa that increase across an environmental gradient from those that decrease to provide a more complete picture of the community response to that gradient.

# 2 Methods

## 2.1 Operation:

Like TITAN2, pTITAN2 was developed using the R programming language (R Core Team 2021). pTITAN2 has been testing on Windows, Mac OS, and Ubuntu for the latest R version (4.1.0 at time of writing) along with old release and development versions via github actions. Users should be familiar with the TITAN2 package operations before using pTITAN2 (see Matthew E. Baker, King, and Kahle (2020)).

The basic workflow for pTITAN2 is

1. Prepare and import the environmental gradient dataset into R
2. Prepare and import the taxonomic dataset into R
3. Preprocess raw taxonomic data to determine appropriate taxonomic level of resloution (occurrence function)
4. Select columns for the taxon level retuned by occurrence function
5. Permut the data across treatment labels to generate list of lists
6. Set up cluster for parallele processing (optional)
7. Run TITAN2 series on original and permuted data sets
8. Analyze probability of exceeding observed diffrence in changepoint between treatments based on distribution of paired changepoint differences

## 2.2 Implementation:

The first step of pTITAN2 is to provide the data about the environmental gradient in exactly the format as for TITAN, step 1). This can be either a single file or include in teh taxonomic data file (Matthew E. Baker, King, and Kahle 2020). Like TITAN2, taxonomic information should be provided as counts or density. Unlike TITAN2, pTITAN2 taxonomic data nees to be provide as a code that is 8 characters in length and captures four levels of hierarchical taxonomic classification information.

The pTITAN2 package provides four example data sets, two taxonomic and two environmental gradient (Table 2.1). These data sets are provided as raw csv files and as prepared R datasets.

Table 1: Table 2.1: Example data sets provided in pTITAN2.

| R Data | csv File | Data Type | Region | Treatment |
| --- | --- | --- | --- | --- |
| C\_IC\_D\_06\_wID | C\_IC\_D\_06\_wID.csv | Environmental Gradient | Chaparral | Dry |
| C\_IC\_N\_06\_wID | C\_IC\_N\_06\_wID.csv | Environmental Gradient | Chaparral | Normal |
| CD\_06\_Mall\_wID | CD\_06\_Mall\_wID.csv | Taxonomic | Chaparral | Dry |
| CN\_06\_Mall\_wID | CN\_06\_Mall\_wID.csv | Taxonomic | Chaparral | Normal |

You can gain access to the csv files via system.file

list.files(system.file("extdata", package = "pTITAN2"))  
## [1] "C\_IC\_D\_06\_wID.csv" "C\_IC\_N\_06\_wID.csv" "CD\_06\_Mall\_wID.csv"  
## [4] "CN\_06\_Mall\_wID.csv"

or get the data sets loaded into your environment via

data(C\_IC\_D\_06\_wID, C\_IC\_N\_06\_wID, CD\_06\_Mall\_wID, CN\_06\_Mall\_wID,  
 package = "pTITAN2")  
  
str(C\_IC\_D\_06\_wID) # Environemntal Gradient, Dry Treatment  
## 'data.frame': 251 obs. of 2 variables:  
## $ StationID: chr "308000" "707063" "707086" "500762" ...  
## $ ImpCover : num 0.0151 0.0907 0.076 3.9944 2.5167 ...  
str(C\_IC\_N\_06\_wID) # Environemntal Gradient, Normal Treatment  
## 'data.frame': 124 obs. of 2 variables:  
## $ StationID: chr "707059" "707076" "707219" "500004" ...  
## $ ImpCover : num 0.05959 0.00695 8.46983 35.8094 12.97293 ...  
dim(CD\_06\_Mall\_wID) # Taxonomic, Dry Treatment  
## [1] 251 501  
dim(CN\_06\_Mall\_wID) # Taxonomic, Normal Treatment  
## [1] 124 501

The CN\_06\_Mall.csv (Chaparral Region, Treatment = Normal) file contains raw macroinvertebrate density data for 500 possible macroinvertebrate codes for each taxonomic level (class, order, family, genus). The occurrences function selects the codes that should be used for the TITAN2::titan run. The goal is to select the macroinvertebrate code with the most taxonomic detail having at least n occurrences. Only one macroinvertebrate code will be associated with the macroinvertebrate counts. For example, if there are at least six occurrences at the genus level, the family, order, and class codes would not be used in the TITAN2::titan run.

The names within the data set are expected to have the following structure:

* 8 characters in length
* characters 1 and 2 denote the class
* characters 3 and 4 denote the order
* characters 5 and 6 denote the family
* characters 7 and 8 denote the genus.

If no information at a level exists, use “00” to hold the place. For example: A code that is ‘Bi000000’ is the Bivalvia class, while BiVe0000 is the Bivalvia class, Veneroida order. BiVeSh00 is the Bivalvia class, Veneroida order, Spheriridae family. BiVeSh01 is a genus within that family.

The first new function provided by pTITAN2 is occurrences. Taking the taxonomic data as an input, the return of occurrences is a data.frame with the taxon, the class, order, family, and genus split out into individual columns, and the count of occurrences within the provided taxonomic data set. TITAN2::titan recommends all taxonomic groups have at least five observations (Matthew E. Baker and King 2010). Thus occurances returns only taxons with at least n observations, defaulting to six. The taxonomic code chosen for analysis should be at the finest possible resolution. For example, if a macroinvertebrate count has at least six occurrences in a genus code, the family, order, and class codes associated with these counts should be removed. Further, if there are too few counts at the genus level, but at least six counts at the family level- the family code would be retained and the order and class codes would be removed.

The second new function provided by pTITAN2 is the permute function which provides a list of permuted sets of taxa and environmental gradients. This function is used with categorical envrionmental variables (treatments), such as Wet/Dry or Urban/Rural. The funciton permutes the treatment labels across the data such that each station has a non-zero probability of being assigned to each treatment, and the stations are unique within each treatment and replication. There are some limitation on the permutations. First, as with TITAN, a site cannot occur in a category more than once within a permutation. Second, the original sample size distribution is maintained. These limitations address potential sample size effects in TITAN, where treatments with low sample sizes have wide confidence intervals and variable change points compared to treatments with high sample sizes, and enable comparisons between treatments with different sample sizes.

For example, assume we have sites A, B, C, D, and E with treatments 1 and 2 (Table 2.2). Let Trt0 denote the initial treatment labels for the sites and Trt1, …, Trt4 denote permuted treatment labels. For sites A and C, each permuted set of treatment labels consist of one row for label 1 and one row for label 2. For sites B, D, and E, the initial observations were for treatments 1, 2, and 2 respectively. The balance of these labels is maintained across the permutations.

Table 2: Table 2.2: Example distribution of sites and perutated treatment labels.

| site | trt0 | trt1 | trt2 | trt3 | trt4 |
| --- | --- | --- | --- | --- | --- |
| A | 1 | 1 | 2 | 1 | 1 |
| A | 2 | 2 | 1 | 2 | 2 |
| B | 1 | 2 | 2 | 2 | 1 |
| C | 1 | 1 | 2 | 2 | 1 |
| C | 2 | 2 | 1 | 1 | 2 |
| D | 2 | 2 | 2 | 1 | 2 |
| E | 2 | 1 | 1 | 2 | 2 |

After permutations, clusters can be used for parallel processing of TITAN::titan() calls. This can be advantageous as TITAN::titan() calls can be time and computationally expensive. Following the needed TITAN::titan() calls the differences between treatment change points in the observed data can be compared to the differences between treatment change points in the permuted stat to determine if the observed treatment differences are statistically significant.

# 3 Example

Here we present an example showing implementation of pTITAN2. We will describe the provided example data sets and how to use the occurrences() and permute() functions.

To reproduce the examples in this vignette you will need to load and attach the pTITAN2 and magrittr namespaces. Other namespaces are used explicitly, loaded (not attached) here.

library(pTITAN2)  
library(magrittr)  
loadNamespace("data.table")  
## <environment: namespace:data.table>  
loadNamespace("dplyr")  
## <environment: namespace:dplyr>  
loadNamespace("tidyr")  
## <environment: namespace:tidyr>

# 11 References

Baker, Matthew E., Ryan S. King, and David Kahle. 2020. *Titan2: Threshold Indicator Taxa Analysis*. <https://CRAN.R-project.org/package=TITAN2>.

Baker, Matthew E, and Ryan S King. 2010. “A New Method for Detecting and Interpreting Biodiversity and Ecological Community Thresholds.” *Methods in Ecology and Evolution* 1 (1): 25–37.

———. 2013. “Of TITAN and Straw Men: An Appeal for Greater Understanding of Community Data.” *Freshwater Science* 32 (2): 489–506.

Dufrêne, Marc, and Pierre Legendre. 1997. “Species Assemblages and Indicator Species: The Need for a Flexible Asymmetrical Approach.” *Ecological Monographs* 67 (3): 345–66.

Fetscher, A Elizabeth, Martha Sutula, Ashmita Sengupta, and Naomi Detenbeck. 2014. *Linking Nutrients to Alterations in Aquatic Life in California Wadeable Streams*. US Environmental Protection Agency, Office of Research; Development ….

Greenland, Sander, Stephen J Senn, Kenneth J Rothman, John B Carlin, Charles Poole, Steven N Goodman, and Douglas G Altman. 2016. “Statistical Tests, p Values, Confidence Intervals, and Power: A Guide to Misinterpretations.” *European Journal of Epidemiology* 31 (4): 337–50. https://doi.org/<https://doi.org/10.1007/s10654-016-0149-3>.

Homer, Collin, Jon Dewitz, Joyce Fry, Michael Coan, Nazmul Hossain, Charles Larson, Nate Herold, et al. 2007. “Completion of the 2001 National Land Cover Database for the Counterminous United States.” *Photogrammetric Engineering and Remote Sensing* 73 (4): 337.

King, Ryan S, Matthew E Baker, Paul F Kazyak, and Donald E Weller. 2011. “How Novel Is Too Novel? Stream Community Thresholds at Exceptionally Low Levels of Catchment Urbanization.” *Ecological Applications* 21 (5): 1659–78.

King, Ryan S, and Curtis J Richardson. 2003. “Integrating Bioassessment and Ecological Risk Assessment: An Approach to Developing Numerical Water-Quality Criteria.” *Environmental Management* 31 (6): 795–809.

Ode, PR, TM Kincaid, T Fleming, and AC Rehn. 2011. “Ecological Condition Assessments of California’s Perennial Wadeable Streams, Highlights from the Surface Water Ambient Monitoring Program’s Perennial Streams Assessment (PSA)(2000–2007).” *A Collaboration Between the State Water Resources Control Board’s Non-Point Source Pollution Control Program (NPS Program), Surface Water Ambient Monitoring Program (SWAMP), California Department of Fish and Game Aquatic Bioassessment Laboratory, and the US Environmental Protection Agency*.

Qian, Song S, Ryan S King, and Curtis J Richardson. 2003. “Two Statistical Methods for the Detection of Environmental Thresholds.” *Ecological Modelling* 166 (1-2): 87–97.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.

Schenker, Nathaniel, and Jane F Gentleman. 2001. “On Judging the Significance of Differences by Examining the Overlap Between Confidence Intervals.” *The American Statistician* 55 (3): 182–86.