Compositional data analysis to model tourists’ expenditure by categories

*work in progress*

Gabriel Brida, Bibiana Lanzilotta, Leonardo Moreno

Facultad de Ciencias Económicas y de Administración
Universidad de la República (Montevideo, Uruguay)

May 18, 2021
Overview

1. Motivation

2. Background

3. Data and Methodology

4. Empirical Results

5. Concluding remarks
Motivation

- How do tourists distribute their expenditure at destination?
- What are the characteristics of the tourist that infer a certain (relative) spending pattern?
- This expenditure pattern has any differential behavior depending on the seasons of the year, the length of stay at destination, etc.?
- Compositional Data Analysis (CoDa) is a statistical tool used when data contain information about the composition of the different parts of the total
- The study models the relative expenditure of tourists among the different items as a function of a set of co-variables by introducing a Dirichlet regression model
On aggregate tourism demand

- There are hundreds of empirical studies using different econometric and statistical methodologies and several published review papers on aggregate tourism demand modeling and forecasting.
- On the other hand, the academic work that analyze the determinants of tourist spending at the microeconomic level has been relatively less and more recent.
- According to [Brida et al, 2018] only few papers analyse the distribution of tourists’ expenditure into the different items (including accommodation, food, transportation, ...). Applications to the study of cruise tourism in Uruguay: [Brida et al, 2020], [Brida et al, 2018]
- And only a bunch uses CoDa to analyse how do tourists distribute their expenditure at destination [Ferrer-Rosell et al, 2015]
Background on CoDa applications and Dirichlet regression

- There is a recent paper that reviews all CoDa applications in tourism and hospitality to date [Coenders & Ferrer-Rosell, 2020]:
- To the best of our knowledge, there are no applications in Tourism of Dirichlet regressions.
- The previous review paper reports only one paper using a Dirichlet regression model and CoDa in the area of Social Sciences [Morais et al, 2018]
The data

- The empirical study considers incoming tourists to Uruguay by means of cross-sections data (2014–2019).
- Individual data comes from the Receptive Tourism Survey in Uruguay -collected by the Ministry of Tourism of Uruguay- which is aimed at non-resident tourists who visit our country (at the country's exit posts).
- The total sample size consists of 70,501 surveys conducted between 2014 and 2019. The analysis will be developed on those tourists who report spending in at least one of the items (69,456 surveys).
- Data is open and available (until 2018 at: https://www.gub.uy/ministerio-turismo/datos-y-estadisticas/microdatos)
- Some descriptive statistics onwards
Methodology: Compositional data in 5 min (1)

- To illustrate, suppose that \( X = (X_1, X_2, X_3) \in \mathbb{R}^3 \) and \( X_i > 0 \) for all \( i = 1, 2, 3 \).
- The vector is normalized with the \( L_1 \) norm, that is,

\[
Y = \left( \frac{X_1}{X_1 + X_2 + X_3}, \frac{X_2}{X_1 + X_2 + X_3}, \frac{X_3}{X_1 + X_2 + X_3} \right) \in S^3,
\]

where \( S^3 = \{ y \in \mathbb{R}^3 : y_1 > 0, y_2 > 0, y_3 > 0, y_1 + y_2 + y_3 = 1 \} \) is the simplex.
Methodology: Compositional data in 5 min (2)

Figure: Transformation of variable X to simplex in $\mathbb{R}^3$. 
Methodology: Compositional data in 5 min (2)

Figure: Simplex representation in dimension 2
Methodology: Compositional data in 5 min (3)

Figure: Parallel coordinates plots (marginal plot)
Methodology: Compositional data in 5 min (4)

Why a particular regression methodology is necessary?
Methodology: Compositional data in 5 min (4)

Why a particular regression methodology is necessary?

- The data are in a linear manifold with boundary.
- In a linear model, the coefficient matrix is singular.
- If \( y \in S^3 \) is the output of a regression model, the assumptions of normality are not possible.
Methodology: Compositional data in 5 min (5)

The Dirichlet Regression

- The density of the Dirichlet distribution (in dimension 3) is
  \[ f_{(\alpha_1,\alpha_2,\alpha_3)}(y_1, y_2, y_3) = \frac{\Gamma(\alpha_1 + \alpha_2 + \alpha_3)}{\alpha_1 \alpha_2 \alpha_3} y_1^{\alpha_1-1} y_2^{\alpha_2-1} y_3^{\alpha_3-1}, \]
  with \( \alpha_i > 0 \) for all \( i = 1, 2, 3 \), \((y_1, y_2, y_3) \in S^3\), and where \( \Gamma \) is the gamma function.

- \( E(Y_i) = \alpha'(i) = \frac{\alpha_i}{\alpha_0} \) and \( V(Y_i) = \frac{\alpha'(i)}{\alpha_0+1} \) with \( \alpha_0 = \alpha_1 + \alpha_2 + \alpha_3 \)

- In the Dirichlet regression the logarithm function is used as a link function in a generalized linear model, that is,
  \[ \log(\alpha_i) = X_i \beta_i, \]
  where \( X_i \) is the vector of covariates and \( \beta_i \) their respective coefficients.
Methodology: Compositional data in 5 min (5)

The Dirichlet Regression

- The density of the Dirichlet distribution (in dimension 3) is

\[
f_{(\alpha_1, \alpha_2, \alpha_3)}(y_1, y_2, y_3) = \frac{\Gamma(\alpha_1 + \alpha_2 + \alpha_3)}{\alpha_1 \alpha_2 \alpha_3} y_1^{\alpha_1-1} y_2^{\alpha_2-1} y_3^{\alpha_3-1},
\]

with \( \alpha_i > 0 \) for all \( i = 1, 2, 3 \), \((y_1, y_2, y_3) \in S^3\), and where \( \Gamma \) is the gamma function.

- \( E(Y_i) = \alpha'(i) = \frac{\alpha_i}{\alpha_0} \) and \( V(Y_i) = \frac{\alpha'(i)}{\alpha_0 + 1} \) with \( \alpha_0 = \alpha_1 + \alpha_2 + \alpha_3 \)

- In the Dirichlet regression the logarithm function is used as a link function in a generalized linear model, that is,

\[
log(\alpha_i) = X_i \beta_i,
\]

where \( X_i \) is the vector of covariates and \( \beta_i \) their respective coefficients.
Dirichlet distribution

**Figure**: Level set Dirichlet’s density and simulation of 200 observations for
Left panel: $\alpha = (0.1, 0.1, 0.1)$. Left center panel: $\alpha = (1, 1, 1)$. Right center panel: $\alpha = (5, 5, 5)$. Right panel: $\alpha = (0.5, 2, 5)$. 
The case study: international tourism in Uruguay (1)

- The distribution of relative expenditure in the different items (including Accommodation (EAcco), foods (EFood) and shopping (EShopp), tours (ETour)) is modeled.
- The following covariates are used:
  - Days of (stay at least one)
  - Nationality (Argentina, Brazil)
  - Destination (Montevideo, Punta del Este)
- $\alpha'$ is a function of the days (for Argentinians and Brazilians).
The case study: international tourism in Uruguay (2)

Figure: $\alpha'(i)$ for tourists who stay between 1 and 20 days. Left panel: Argentinians. Center panel: Brazilians. Right panel: Uruguayans.
The case study: international tourism in Uruguay (2)

Figure: $\alpha'(i)$ for tourists who stay between 1 and 20 days. As destination: Montevideo (left panel), Punta del Este (right panel).
Descriptive statistics

Table: Proportions of answers in each covariate.

<table>
<thead>
<tr>
<th>Nationality</th>
<th>%</th>
<th>Destination</th>
<th>%</th>
<th>Age (years)</th>
<th>%</th>
<th>Profession</th>
<th>%</th>
<th>Accomm.</th>
<th>%</th>
<th>Reason</th>
<th>%</th>
<th>Year</th>
<th>%</th>
<th>Month</th>
<th>%</th>
<th>Days</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uruguay</td>
<td>12.8</td>
<td>Montevideo</td>
<td>37.9</td>
<td>[0,20]</td>
<td>1.2</td>
<td>Public Worker</td>
<td>27.3</td>
<td>Apart Hotel</td>
<td>2.5</td>
<td>Family view</td>
<td>20.4</td>
<td></td>
<td></td>
<td>12.2</td>
<td></td>
<td>0</td>
<td>5.5</td>
</tr>
<tr>
<td>Argentina</td>
<td>57.2</td>
<td>Pta del Este and nearby</td>
<td>29.7</td>
<td>[20,40]</td>
<td>40.3</td>
<td>Professional</td>
<td>27.8</td>
<td>Camping Hostel</td>
<td>4.6</td>
<td>Transit</td>
<td>4.4</td>
<td></td>
<td></td>
<td>10.0</td>
<td>1-3</td>
<td>34.2</td>
<td></td>
</tr>
<tr>
<td>Brazil</td>
<td>13.2</td>
<td>Colonia</td>
<td>9.4</td>
<td>[40,60]</td>
<td>53.6</td>
<td>Businessman</td>
<td>16.5</td>
<td>Hotel</td>
<td>41.1</td>
<td>Holidays</td>
<td>57.2</td>
<td></td>
<td></td>
<td>8.6</td>
<td>4-7</td>
<td>39.1</td>
<td></td>
</tr>
<tr>
<td>Paraguay</td>
<td>1.3</td>
<td>Thermal</td>
<td>14.0</td>
<td>[60,100]</td>
<td>4.9</td>
<td>Retired</td>
<td>7.8</td>
<td>None</td>
<td>5.4</td>
<td>Studies</td>
<td>0.7</td>
<td>2017</td>
<td>19.6</td>
<td>4</td>
<td>7.2</td>
<td>8-14</td>
<td>17.3</td>
</tr>
<tr>
<td>Chile</td>
<td>1.9</td>
<td>Transit</td>
<td>4.4</td>
<td></td>
<td>2.9</td>
<td>Leader</td>
<td>6.3</td>
<td>Others</td>
<td>0.9</td>
<td>2nd Residence</td>
<td>3.1</td>
<td>2018</td>
<td>18.3</td>
<td>5</td>
<td>6.2</td>
<td>15 or +</td>
<td>3.9</td>
</tr>
<tr>
<td>USA</td>
<td>3.3</td>
<td>Others</td>
<td>4.7</td>
<td></td>
<td>2.9</td>
<td>Visitor relatives</td>
<td>6.9</td>
<td>Religious</td>
<td>0.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rest of America</td>
<td></td>
<td>Student</td>
<td></td>
<td></td>
<td>3.8</td>
<td>Rental House</td>
<td>6.9</td>
<td>Religious</td>
<td>0.3</td>
<td></td>
<td></td>
<td>7</td>
<td>7.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Europe</td>
<td>5.6</td>
<td>Boss</td>
<td></td>
<td></td>
<td>4.0</td>
<td>Second home</td>
<td>5.9</td>
<td>Health</td>
<td>0.5</td>
<td></td>
<td></td>
<td>8</td>
<td>6.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Others</td>
<td>1.0</td>
<td>Worker</td>
<td></td>
<td></td>
<td>0.8</td>
<td>Business</td>
<td>11.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>9</td>
<td>6.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Athlete</td>
<td></td>
<td>Others</td>
<td>1.8</td>
<td></td>
<td>1.0</td>
<td>Others</td>
<td>1.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10</td>
<td>7.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Others</td>
<td></td>
<td></td>
<td>1.8</td>
<td></td>
<td>1.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>11</td>
<td>8.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Brida-Lanzilotta-Moreno (FCEA, UdelaR) Compositional data analysis to model tourists May 18, 2021
Descriptive statistics

**Figure:** Average of the proportions of tourists according to their origin, by year and season.
Descriptive statistics
Descriptive statistics

Figure: parallel coordinates plots: Nationality

Brida-Lanzilotta-Moreno (FCEA, UdelaR)  Compositional data analysis to model tourists  May 18, 2021
Empirical Results

Descriptive statistics
Dirichlet models (\(\beta\) coefficients)

Table: Model 1: with covariates Season and Year

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficients</th>
<th>Accommodation</th>
<th>Expenditure</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>Std. Error</td>
<td>Estimate</td>
<td>Std. Error</td>
</tr>
<tr>
<td>Intercept</td>
<td>-1.547</td>
<td>0.00075</td>
<td>-0.888</td>
<td>0.00082</td>
</tr>
<tr>
<td>Season</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Winter</td>
<td>-0.133</td>
<td>0.00059</td>
<td>0.235</td>
<td>0.00064</td>
</tr>
<tr>
<td>Spring/autumn</td>
<td>-0.76</td>
<td>0.00059</td>
<td>0.216</td>
<td>0.00064</td>
</tr>
<tr>
<td>2015</td>
<td>0.012</td>
<td>0.00086</td>
<td>-0.077</td>
<td>0.00093</td>
</tr>
<tr>
<td>2016</td>
<td>0.101</td>
<td>0.00086</td>
<td>0.097</td>
<td>0.00091</td>
</tr>
<tr>
<td>Year</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2017</td>
<td>0.205</td>
<td>0.00081</td>
<td>0.153</td>
<td>0.00088</td>
</tr>
<tr>
<td>2018</td>
<td>0.141</td>
<td>0.00082</td>
<td>0.213</td>
<td>0.00089</td>
</tr>
<tr>
<td>2019</td>
<td>0.137</td>
<td>0.00085</td>
<td>0.346</td>
<td>0.00093</td>
</tr>
</tbody>
</table>

Interpretation: \(\hat{\alpha} = (e^{-1.547+0.205}, e^{-0.888+0.153}, e^{-0.556+0.211})\)
(High-2017) \(\sum \hat{\alpha}_i = (0.18, 0.33, 0.49)\)
(High-2019) \((0.15, 0.38, 0.47)\)
Dirichlet models ($\beta$ coefficients)

Table: Model 2: with covariates Destination, Accommodation and Nationality

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficients</th>
<th>Expenditure</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Accommodation</td>
<td>Food</td>
<td>Others</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Estimate</td>
<td>Std. Error</td>
<td>Estimate</td>
<td>Std. Error</td>
<td>Estimate</td>
<td>Std. Error</td>
</tr>
<tr>
<td>Intercept</td>
<td>1.059</td>
<td>0.0006</td>
<td>0.469</td>
<td>0.0006</td>
<td>0.85</td>
<td>0.0006</td>
</tr>
<tr>
<td>Pta del Este and nearby</td>
<td>0.081</td>
<td>0.0006</td>
<td>0.226</td>
<td>0.0007</td>
<td>0.147</td>
<td>0.0007</td>
</tr>
<tr>
<td>Thermal</td>
<td>0.15</td>
<td>0.0007</td>
<td>0.227</td>
<td>0.0008</td>
<td>0.118</td>
<td>0.0008</td>
</tr>
<tr>
<td>Colonia</td>
<td>-0.276</td>
<td>0.0009</td>
<td>-0.166</td>
<td>0.0009</td>
<td>-0.628</td>
<td>0.001</td>
</tr>
<tr>
<td>Others</td>
<td>-0.125</td>
<td>0.0011</td>
<td>-0.082</td>
<td>0.0012</td>
<td>-0.039</td>
<td>0.0013</td>
</tr>
<tr>
<td>Transit</td>
<td>-0.23</td>
<td>0.0014</td>
<td>-1.038</td>
<td>0.0016</td>
<td>-0.061</td>
<td>0.0015</td>
</tr>
</tbody>
</table>
Dirichlet models ($\beta$ coefficients)

Table: Model 2: with covariates Destination, Accommodation and Nationality

<table>
<thead>
<tr>
<th>Accommodation</th>
<th>Camp-Hostel</th>
<th>Cottage</th>
<th>None</th>
<th>Others</th>
<th>Visiting relatives</th>
<th>Rental house</th>
<th>Second home</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.98</td>
<td>-0.079</td>
<td>-3.364</td>
<td>-3.21</td>
<td>-3.41</td>
<td>0.01</td>
<td>-0.05</td>
</tr>
<tr>
<td></td>
<td>0.0012</td>
<td>0.0014</td>
<td>0.0014</td>
<td>0.0025</td>
<td>0.0006</td>
<td>0.0008</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>-0.235</td>
<td>0.149</td>
<td>-1.369</td>
<td>-0.868</td>
<td>-0.556</td>
<td>0.133</td>
<td>0.028</td>
</tr>
<tr>
<td></td>
<td>0.0012</td>
<td>0.0014</td>
<td>0.0016</td>
<td>0.0029</td>
<td>0.0007</td>
<td>0.0008</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>-0.487</td>
<td>-0.197</td>
<td>-1.392</td>
<td>-0.818</td>
<td>-0.286</td>
<td>0.051</td>
<td>0.068</td>
</tr>
<tr>
<td></td>
<td>0.0012</td>
<td>0.0014</td>
<td>0.0015</td>
<td>0.0031</td>
<td>0.0008</td>
<td>0.0009</td>
<td>0.001</td>
</tr>
</tbody>
</table>
### Dirichlet models ($\beta$ coefficients)

**Table:** Model 2: with covariates Destination, Accommodation and Nationality

<table>
<thead>
<tr>
<th>Nationality</th>
<th>Uruguay</th>
<th>0.008</th>
<th>-0.269</th>
<th>0.0009</th>
<th>-0.1</th>
<th>0.001</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Brazil</td>
<td>-0.388</td>
<td>-0.409</td>
<td>0.0007</td>
<td>-0.734</td>
<td>0.0008</td>
</tr>
<tr>
<td></td>
<td>Paraguay</td>
<td>-0.024</td>
<td>-0.035</td>
<td>0.002</td>
<td>0.008</td>
<td>0.0021</td>
</tr>
<tr>
<td></td>
<td>Chile</td>
<td>-0.1</td>
<td>-0.095</td>
<td>0.0018</td>
<td>-0.117</td>
<td>0.0018</td>
</tr>
<tr>
<td></td>
<td>N. America</td>
<td>0.001</td>
<td>0.0015</td>
<td>-0.04</td>
<td>0.0016</td>
<td>-0.143</td>
</tr>
<tr>
<td></td>
<td>Other Amer.</td>
<td>-0.001</td>
<td>0.0014</td>
<td>0.022</td>
<td>0.0014</td>
<td>-0.093</td>
</tr>
<tr>
<td></td>
<td>Europe</td>
<td>-0.022</td>
<td>-0.003</td>
<td>0.0012</td>
<td>-0.069</td>
<td>0.0012</td>
</tr>
<tr>
<td></td>
<td>Others</td>
<td>-0.058</td>
<td>-0.05</td>
<td>0.063</td>
<td>-0.402</td>
<td>0.0027</td>
</tr>
</tbody>
</table>
Other methodologies

There are other ways to analyze this type of data. For example, you can use a transformation in a space of less dimension (logratio transformation), see for instance [Filzmoser et al, 2018].

The most commonly used is the logratio transformation, see [Aitchison, 1982]

Some difficulties of this model are,

- “This can give greater importance to components that have little overall emphasis on a meaningful understanding of the composition...[Hijazi & Jernigan, 2009].”
- “…the resulting parameters are only interpretable in the transformed space and have no straightforward meaning.[Maier, 2014].”
- “if heteroscedasticity persists after the transformation, one has to either violate the assumption of homoscedasticity in linear models, or incorporate model terms capturing heteroscedasticity that further complicate the model (and interpretation).[Maier, 2014].”
Problem of zeros

(Problem of zeros) The use of the Dirichlet regression implies that the data set should consist of strictly positive values, but this data set contain zeros. One strategy used to solve this problem is to replace the zeros with small positive values, see [Greenacre, 2018]. We apply the transformation \( y^* = \frac{y(n-1)+1/d}{n} \) where \( n \) and \( d \) are the sample size and the space dimension respectively.

On the other hand, in [Tsagris & Stewart, 2018], a modification in the Dirichlet regression is proposed to incorporate the existence of zeros (where it is not necessary to transform them into positive values).
Empirical Results

A brief summary of the results

- From compositional analysis
  - Accommodation and tours decreases participation on expenditure on food and shopping increases share, as tourists stay longer.
  - This phenomenon is diverse for both Argentine and Brazilian tourists (with the non-residents Uruguayans, the principal sources of tourists for Uruguay).
  - In Brazilian tourists, the increase of the share of expenditure on food and shopping and the decrease of accommodation and tours as the stay extended, is more pronounced..
  - When the stay is less, the dispersion reduces
  - As expected, the type of accommodation affects the pattern of tourist expenditure and the spending on accommodation..
A brief summary of the results

- From Dirichlet model: what information does it add to the CoDa analysis?
  - (Model 1) Concentration on accommodation diminishes in low and middle tourist seasons and food and others increase in middle season.
  - (Model 1) Spending on accommodation seems to reduce participation since 2017, while food and other expenditures increase.
  - (Model 2): When the tourist destination is neither Montevideo nor in the sun & beach destinations, concentration in accommodation also reduces.
  - (Model 2): The pattern of spending also changes according to nationality. The concentration of expenditure on accommodation is higher in the case of Argentines and North Americans. Alternatively, the concentration of food is higher in Brazilian tourists than in other nationalities.
Concluding remarks

- Two models are analysed (all variables are significant in both models):
  1. including only tourism season and year;
  2. including also tourist destination, type of accommodation and nationality.

- Expenditure pattern in accommodation, food and others varies (significantly) according to the tourist season, the tourist destination, the type of accommodation and nationality.
There are significant differences in the proportion of spending on accommodation, food and shopping (or other expenses) when tourism does not take place in high season.

Additionally, when the tourist destination is neither Montevideo nor in the sun & beach destinations, concentration in accommodation also reduces.

The study brings tools to analyze the behavior of tourists as consumers visiting Uruguay as related to expenditure pattern at microeconomic level. Particularly useful for doing adjustments on the tourism supply according to the demand behaviour.
Agenda

- Extend the analysis to previous years using the CoDa and Dirichlet regressions and analyze the dynamics of the problem.
- Introducing alternative models (to substitute Dirichlet regression). Seemingly Unrelated Regression (SUR) methodology?
- Consider additional variables: origin of tourists, gender of type of tour group.
- Strengthen the analysis of the model’s residuals to get all the information they contain.
- Study of domestic tourism, given its foreseeable importance in (future) post-covid times.
- Replicate the study for other destinations for comparing results and analyse the robustness of the methodology.
References I


References II


Thank you!