Principal component analysis for functional data on grain yield of winter wheat cultivars

Miroslaw Krzyisko¹, Adriana Derejko², Tomasz Gorecki¹, Edward Gacek³

¹Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Umultowska 87, 61-614 Poznan, Poland, e-mail: mkrzysko@amu.edu.pl, tomasz.gorecki@amu.edu.pl
²Department of Experimental Design and Bioinformatics, Warsaw University of Life Sciences, Poland, email: adrainaderejko@gmail.com
³The Research Centre for Cultivar Testing, Slupia Wielka, Poland, email: e.gacek@coboru.pl

SUMMARY

The aim of this paper is to present a statistical methodology to assess patterns of cultivars’ adaptive response to agricultural environments (agro-ecosystems) on the basis of complete Genotype x Crop Management x Location x Year (GxMxLxY) data obtained from 3-year multi-location two-factor trials conducted within the framework of the Polish post-registration trials (PDOiR), with an illustration of the application and usefulness of this methodology in analyzing winter wheat grain yield. Producing specific varieties for each subregion of a target region, from widely adapted varieties, may exploit positive genotype x location (GL) interactions to increase crop yields. Experiments designed to examine combinations of environment (E), management practices (M) and cultivars (G) also provide evidence of the relative importance of each of these factors for yield improvement. The evidence shows that variation due to E far outweighs the variation of grain yield that can be attributed to M or G, or the interactions between these factors, and between these factors and E (Anderson, 2010). This statistical method involves the use of functional PCA and cluster analysis.

A total of 24 cultivars were evaluated over 3 years in 20 environments using randomized incomplete split-block designs with two replications per trial. The methodology proved an efficient tool for the reliable classification of 24 winter wheat cultivars, distinguishing cultivar groups that exhibited homogeneous adaptive response to environments. It enables the identification of cultivars displaying wide or specific adaptation. The remaining cultivars were locally adapted to some testing environments, or some of them were not relatively adapted to the environments because they always yielded substantially below the environmental means. Performing earlier specific selection, or adopting distinct genetic bases for each agro-ecosystem, may further increase the advantage of specific breeding.

Key words: cultivar adaptation, functional data, grain yield, post-registration trials, principal component analysis, winter wheat
1. Introduction

Wheat is a major crop contributing to the world’s food supply. Of the total supply of wheat, an average of 53% is consumed as food in the developed countries, and close to 85% in the developing countries (Denčić et al., 2011). It has long been recognized that wheat yields and other agronomic and quality traits vary considerably as a result of genotype, environment and their interaction (Allard and Bradshaw, 1964; Basford and Cooper, 1998; Trethowan and Crossa, 2007; Denčić et al., 2011).

The main objective of plant breeding in major crop species, including winter wheat, is to develop new cultivars demonstrating one of two adaptation patterns, called wide or specific (local) adaptation, to environments within a target production (growing) area (Sivapalan et al., 2000; Annicchiarico, 2002a,b; Rane et al., 2007). The adaptation patterns of each tested cultivar can be described by their yield responses (also called cultivar adaptive responses – Annicchiarico et al., 2006b, 2011; Annicchiarico and Iannucci, 2008; Gauch et al., 2008) across a wide range of environments and also years in the production area. Predicting repeatable cultivar adaptive responses requires the conduct of 3-year multi-environment trials with a set of offered cultivars, continued across representative test environments (locations) of the production area and years (Annicchiarico, 2002a,b; Trethowan and Crossa, 2007; Annicchiarico et al, 2010). On the basis of yield data from these trials it is possible to estimate (predict) both the genotypic mean yield and repeatable genotype x location (GL) interaction effects (Yan and Hunt, 1998; Trethowan et al., 2002; Annicchiarico, 2002b; Annicchiarico et al. 2006a, 2010). GxL interaction effects for a given cultivar produce repeatable cultivar adaptive responses (Ghaderi et al., 1982; Yan et al., 2007; Rodriguez et al., 2008; Annicchiarico, 2002b).

Cultivars having wide adaptation are defined as those that produce yields substantially above the environmental means and are among the few top-ranking cultivars at a majority of locations across a production area characterized by substantial variation in environmental conditions (Braun et al., 1996; Annicchiarico, 2002b; Rodriguez et al., 2008). Such cultivars produce relatively high and stable yields within the area (Annicchiarico, 2002b; Singh et al., 2007; Yan et al., 2007; Yang et al., 2009). Cultivars having specific adaptation are defined as those that produce yields substantially above the environmental means and are among the few top-ranking cultivars in the range of a sub-region (macroenvironment) within the target region, usually of limited environmental variation (Gauch and Zobel, 1997;
Annicchiarico, 2002b; Lillemo et al., 2005; de la Vega and Chapman, 2006) or in at least one environment within the target area (Annicchiarico and Iannucci, 2008; Annicchiarico et al., 2010). Usually, cultivars with wide adaptation have fairly high yield potential and stress tolerance, whereas specifically-adapted ones have maximum levels of either yield potential or stress tolerance (Annicchiarico, 2002b; Singh et al., 2007; Trethowan and Crossa, 2007; Ulukan, 2008). Although widely adapted cultivars are usually preferred, the merits of those with local adaptation are also recognized (Annicchiarico, 2002a,b; Zhang et al., 2006; Singh et al., 2007).

In Poland, predictions of repeatable cultivar adaptive responses of important crops are assessed in post-registration variety testing trials (called PDOiR trials), which are repeated across environments and years. These deliver essential information for effective cultivar recommendations for megaenvironments (locally adapted cultivars) or for large sub-regions including also the whole country (widely adapted cultivars). Many different statistical methods have been used to estimate or predict cultivar adaptive responses using PDOiR data, including largely graphical methods (Yan et al., 2007; Yang et al., 2009; Kozak, 2010a,b) and more advanced methods (Gauch 1992, 2006; Gauch and Zobel, 1997; Annicchiarico, 2002b; Gauch et al., 2008). Among these are advanced methods based on functional principal components.

However, in a case with a large number of assessed cultivars, the tool of nominal yield used in its classic form can be less effective, due to the many lines on the nominal yield graph and difficulties in clearly distinguishing them (Haussmann et al., 2000; Kozak, 2010b). A solution to this problem could be to group cultivar adaptive responses into homogeneous groups using cluster analysis.

The objective of this study is to illustrate the application and usefulness of functional principal components and cluster analysis to assess the genotype grain yield adaptive responses of recent Polish and foreign winter wheat cultivars across a major wheat-growing area in Poland under integrated crop management, using data obtained in the PDOiR trials.

2. Materials and methods

2.1. Experimental material

The data used in the study relate to the grain yield of 24 recent Polish and foreign winter wheat cultivars tested across 20 locations (called Exp-
imental Stations for Cultivar Testing) and repeated over three growing years (2007–2009). The cultivars were assessed under the post-registration variety testing trials (PDOiR trials) conducted within the nationwide PDOiR trials system developed by the Research Centre for Cultivar Testing in Ślupia Wielka, near Poznań, Poland (http://www.coboru.pl/English/index_eng.aspx). The test locations were selected in such a way as to cover (represent) the main Polish wheat-growing area. In each trial, integrated crop management was used with N-rates of 40 kg ha\(^{-1}\) less as compared with yield expectations and standard PK fertilizations for a given location and with pesticide use limited to seed treatment, without the use of growth regulators to prevent lodging. The seeding rate ranged from 450 to 550 grains/m\(^2\) at the locations, depending on the cultivar. All experiments in macroenvironments were designed as a randomized incomplete split-block design with two replicates, with plot size 15 m\(^2\) (10 × 1.5m). Mean values for varieties are calculated from single experiments. These experiments are incomplete, hence the mean values are correlated. This fact is ignored.

The cultivars, the breeding companies which produced them, and the years of their release are given in Table 1. Among the 24 tested cultivars, 15 come from Polish breeding companies, and the remaining 9 from German, French and UK companies. The locations of the Experimental Stations for Cultivar Testing (SDOOs), whose names and geographical position are indicated in Figure 1, were well-scattered across the main Polish common wheat-growing area, and are thus representative of that area.

### 2.2. Principal component analysis for functional data

For a given variety of winter wheat and a given location (experimental station), the grain yield data are recorded in the form of the following sequence:

1. first intensity of crop management, in 2007
2. second intensity of crop management, in 2007
3. first intensity of crop management, in 2008
4. second intensity of crop management, in 2008
5. first intensity of crop management, in 2009
Table 1. Winter wheat cultivars tested in post-registration trials (PDOiR trials) carried out at locations in 2007–2009

<table>
<thead>
<tr>
<th>Cultivar</th>
<th>Year of release</th>
<th>Breeding company</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alcazar</td>
<td>2006</td>
<td>Secobra Recherches</td>
</tr>
<tr>
<td>Anthus</td>
<td>2002</td>
<td>KWS Lochow GmbH</td>
</tr>
<tr>
<td>Batuta</td>
<td>2003</td>
<td>DANKO Hodowla Roślin sp. z o.o.</td>
</tr>
<tr>
<td>Bogatka</td>
<td>2001</td>
<td>DANKO Hodowla Roślin sp. z o.o.</td>
</tr>
<tr>
<td>Boomer</td>
<td>2003</td>
<td>RAGT Seeds Ltd.</td>
</tr>
<tr>
<td>Dorota</td>
<td>2001</td>
<td>RAGT Seeds Ltd.</td>
</tr>
<tr>
<td>Flair</td>
<td>1998</td>
<td>Saatzucht Hans Schweiger and Co. oHG</td>
</tr>
<tr>
<td>Kobiera</td>
<td>1999</td>
<td>Hodowla Roślin Rośniczych - Nabiona Kobierzyc</td>
</tr>
<tr>
<td>Kris</td>
<td>1997</td>
<td>RAGT Seeds Ltd.</td>
</tr>
<tr>
<td>Legenda</td>
<td>2002</td>
<td>Poznańska Hodowla Roślin sp. z o.o.</td>
</tr>
<tr>
<td>Ludwig</td>
<td>2001</td>
<td>Saatzucht Donau Ges.m.b.H. and CoKG</td>
</tr>
<tr>
<td>Mewa</td>
<td>1996</td>
<td>DANKO Hodowla Roślin sp. z o.o.</td>
</tr>
<tr>
<td>Nadoyna</td>
<td>1999</td>
<td>Poznańska Hodowla Roślin sp. z o.o.</td>
</tr>
<tr>
<td>Nutka</td>
<td>1998</td>
<td>Hodowla Roślin Strzelce sp. z o.o. Grupa IHAR</td>
</tr>
<tr>
<td>Rapsodia</td>
<td>2000</td>
<td>RAGT Seeds Ltd.</td>
</tr>
<tr>
<td>Rywalka</td>
<td>1999</td>
<td>Hodowla Roślin Strzelce sp. z o.o. Grupa IHAR</td>
</tr>
<tr>
<td>Satyna</td>
<td>2000</td>
<td>Hodowla Roślin Rośniczych - Nasiona Kobierzyc</td>
</tr>
<tr>
<td>Smuga</td>
<td>2001</td>
<td>DANKO Hodowla Roślin sp. z o.o.</td>
</tr>
<tr>
<td>Sukces</td>
<td>1996</td>
<td>Hodowla Roślin Strzelce sp. z o.o. Grupa IHAR</td>
</tr>
<tr>
<td>Tonacja</td>
<td>1998</td>
<td>Hodowla Roślin Strzelce sp. z o.o. Grupa IHAR</td>
</tr>
<tr>
<td>Trend</td>
<td>2000</td>
<td>KWS Lochow GmbH</td>
</tr>
<tr>
<td>Turnia</td>
<td>1998</td>
<td>Małopolska Hodowla Roślin-HBP sp. z o.o.</td>
</tr>
<tr>
<td>Wydmya</td>
<td>2002</td>
<td>Hodowla Roślin Smolice sp. z o.o. Grupa IHAR</td>
</tr>
<tr>
<td>Zryta</td>
<td>1996</td>
<td>Hodowla Roślin Strzelce sp. z o.o. Grupa IHAR</td>
</tr>
</tbody>
</table>

This sequence is then supplemented with the data relating to other locations. In this way the given winter wheat variety is characterized by a spatial and temporal sequence composed of \( J = 6 \times 20 = 120 \) numbers.

The successive values of this sequence will be \( y_j \), where \( j = 1, 2, \ldots, J \). The sequence is smoothed using a continuous function \( x(t) \), where \( t \in [0, T] \). The function \( x(t) \) is taken to be of the form

\[
x(t) = \sum_{k=0}^{K} c_k \varphi_k(t), \quad t \in [0, T],
\]

where \( \{ \varphi_k(t) \} \) is an orthonormal system of basis functions, that is, functions for which

\[
\int_0^T \varphi_k(t) \varphi_l(t) dt = \begin{cases} 
1, & \text{if } k = l, \\
0, & \text{if } k \neq l.
\end{cases}
\]
Figure 1. Locations of the Experimental Stations for Cultivar Testing in Poland within the network of COBORU stations where post-registration trials (PDOiR trials) for winter wheat were carried out in 2007-2009.

As an orthonormal basis of $L_2(I)$ we took the orthonormal system of Legendre polynomials (Abramowitz and Stegun, 1965) in the space $L_2([-1,1])$:

$$\tilde{p}_k(x) = \sqrt{\frac{2k+1}{2}} p_k(x),$$

where

$$p_{k+1}(x) = \frac{1}{k+1} [(2k+1)xp_k(x) - kp_{k-1}(x)], \quad k \geq 1,$$

$$p_0(x) = 1, \quad p_1(x) = x.$$

The coefficients $c_0, c_1, \ldots, c_K$ are selected by the least squares method based on the data $y_1, y_2, \ldots, y_J$, i.e. so as to minimize the function:

$$S(c_0, c_1, \ldots, c_K) = \sum_{j=1}^{J} (y_j - \sum_{k=0}^{K} c_k \varphi_k(t_j))^2.$$
In matrix notation, the function $S$ has the form:

$$S(c) = (y - \Phi c)'(y - \Phi c),$$

where $y = (y_1, y_2, \ldots, y_J)'$, $c = (c_0, c_1, \ldots, c_K)'$ and $\Phi$ is a $J \times (K + 1)$ matrix containing the values $\varphi_k(t_j)$, where $k = 0, 1, \ldots, K$, $j = 1, 2, \ldots, J$.

By differentiating $S(c)$ with respect to the vector $c$ we obtain a system of normal equations in the form:

$$2\Phi'\Phi c - 2\Phi'y = 0.$$

Hence the estimate of vector $c$ is equal to:

$$\hat{c} = (\Phi'\Phi)^{-1}\Phi'y.$$

In this way, each of the 24 varieties of winter wheat was characterized by a continuous function

$$x(t) = \sum_{k=0}^{K} \hat{c}_k \varphi_k(t), \; t \in [0,T].$$

The functions $x_i(t)$ are called functional data (Ramsay and Silverman, 2005). The smoothness degree of the function $x(t)$ depends on $K$, since small (large) values of $K$ induce more (less) smoothed curves. The optimal number $K$ of basis elements is determined using the Bayesian information criterion (BIC) for each function $x_i(t)$ separately:

$$\text{BIC} = \ln(\frac{e'e}{2}) + (K + 1)\frac{\ln J}{J},$$

where $e = (e_1, e_2, \ldots, e_J)'$, $e_j = y_j - \sum_{k=0}^{K} \hat{c}_k \varphi_k(t_j)$, $j = 1, 2, \ldots, J$. As the optimum value of $K$ we take the value for which the BIC attains a minimum. Next, among the 24 values of $K$ corresponding to all of the winter wheat cultivars, the modal value is selected, as a common value of $K$ for all 24 functions $x(t)$.

The estimated values $\hat{c}_0, \hat{c}_1, \ldots, \hat{c}_K$ are treated as the values of new statistical properties. For the properties so obtained, we construct principal components. Principal component analysis is based on evaluation of a covariance matrix $\Sigma$ built from the data for $n = 24$ winter wheat cultivars. Let $\hat{C}$ be a matrix in the form
\[ \hat{C} = \begin{bmatrix}
\hat{c}_{10} & \hat{c}_{11} & \ldots & \hat{c}_{1K} \\
\hat{c}_{20} & \hat{c}_{21} & \ldots & \hat{c}_{2K} \\
\vdots & \vdots & \ddots & \vdots \\
\hat{c}_{n0} & \hat{c}_{n1} & \ldots & \hat{c}_{nK}
\end{bmatrix}, \]

where \( \hat{c}_{rs} \) is the \( s \)th coefficient for the \( r \)th cultivar, \( r = 1, 2, \ldots, 24 \), \( s = 1, 2, \ldots \)

Then the estimator \( \hat{\Sigma} \) of the unknown matrix \( \Sigma \) has the following form:

\[ \hat{\Sigma} = \frac{1}{n} \hat{C}'\hat{C}. \]

We next find the eigenvalues \( \lambda_k \) and corresponding eigenvectors \( u_k \) of the matrix \( \hat{\Sigma} \). The pair \((\lambda_k, u_k)\) will be called the \( k \)th principal configuration. Knowing the \( k \)th principal configuration for the random vector \( \mathbf{c} = (c_0, c_1, \ldots, c_K)' \), we are able to find the principal components for the stochastic process \( X(t) \) by determining the \( k \)th principal configuration \((\sigma_k, u_k(t))\) for that process.

**Theorem 1.** (see Górecki and Krzyśko, 2012) The \( k \)th principal configuration of a random vector \( \mathbf{c} \), defined by \((\lambda_k, u_k)\), is related to the \( k \)th principal configuration of the stochastic process \( X(t) \), \((\sigma_k, u_k(t))\), as follows:

\[ \sigma_k = \lambda_k, \quad u_k(t) = u_k'\varphi(t). \]

The functions \( u_k(t) \) are called eigenfunctions or weight functions. The \( k \)th functional principal component for the process \( X(t) \) is defined as the scalar product

\[ \langle u_k(t), x(t) \rangle = \int_0^T u_k(t)x(t)dt. \]

Hence the coordinate projection of the \( i \)th realization \( x_i(t) \) of the process on the \( k \)th functional principal component is equal to

\[ U_{ik} = \hat{c}_i u_k, \]

where

\[ \hat{c}_i = (\hat{c}_{i0}, \hat{c}_{i1}, \ldots, \hat{c}_{iK})', \quad \hat{u}_k = (\hat{u}_{k0}, \hat{u}_{k1}, \ldots, \hat{u}_{kK})', \]

\( i = 1, 2, \ldots, n, \quad k = 1, 2, \ldots \)
The most frequently considered objects are presented on a plot of the first two functional principal components. In this case the criterion of goodness of the constructed functional principal components is the following expression (degree of explanation of the variation of winter wheat yields in the system of the first two functional principal components):

\[
\frac{\lambda_1 + \lambda_2}{\sum \lambda_i} \times 100\% ,
\]

The greater the value of the expression, the greater is the variability shown by the first two functional principal components. In fact, the degree of explanation of the variation using the first two functional principal components relates to the functions \(x_i(t)\) and not the varieties that they represent. The degree of explanation of the variation for the varieties may be smaller, because it depends on the accuracy with which the real data for the varieties are expressed by these functions.

In the computational process the R program (R Core Team, 2013) was used. We also used the fda package (Ramsay et al., 2012).
3. Results

Figure 2 shows a scree graph for the functional principal component analysis. It can be seen that the first functional principal component explains 67.49% of the variation, and the second explains 26.64%. In total, the first two functional principal components explain as much as 94.13% of the variation. In this situation, a projection onto a plane constructed using the first two functional principal components would appear to be a very good solution. The algorithm selected 4 \((K = 3)\) elements of the orthonormal basis of Legendre polynomials.

Figure 3 shows the first two weight functions \(u_k(t), \ k = 1, 2\) for the 4 elements of the orthonormal basis of Legendre polynomials.

Finally there remains the most important visualization, namely the projection of the data onto the plane of the first two functional principal components, which will show which cultivars of winter wheat behaved similarly during the period considered, in terms of the tested properties. If the points lie close to each other, this means that the cultivars behaved similarly in this regard. The graph is shown in Figure 4.

Now, in order to form homogeneous groups of cultivars, we apply cluster analysis. In the first step, we construct the minimum spanning tree. Two
Figure 4. Projection on the first two functional principal components. Homogeneous groups contained in the ellipses

Popular algorithms for constructing the minimum spanning tree are those of Kruskal (1956) and Prim (1957). Kruskal’s algorithm is a greedy algorithm that adds edges in order of their length, so long as they do not result in a cycle. Prim’s algorithm starts with a tree consisting of a single vertex, and keeps adding the smallest edge of the tree. The minimum spanning tree can be used for clustering. This idea is to break (remove edges from) the minimum spanning tree at edges that are greater than the common critical value $\rho^*$. We compute the mean $\bar{x}$ and the standard deviation $s$ of all edge lengths in the minimum spanning tree, and eliminate edges which are longer than

$$\rho^* = \bar{x} + s,$$

(when we used $2s$ instead of $s$ we obtained only one large cluster). The points which remain connected in the minimum spanning tree form pseudo-clusters (Karoński, 1973). The distance between two cultivars $i$ and $j$ is
equal to the Euclidean distance of the form

\[ d(S_i, S_j) = \sqrt{\sum_{k=1}^{4} (S_i^{(k)} - S_j^{(k)})^2}, \]

where the \( i \)th cultivar is represented by the vector \( S_i = (S_i^{(1)}, \ldots, S_i^{(4)})' \) of functional principal components. In Figure 4 are plotted the values of the first two functional principal components, with homogeneous groups outlined in the ellipses. The resulting functional principal components can also be called the interactive principal components. Included in them is in fact the adaptive response of the grain yield of 24 varieties to the environment, years and intensity of crop management. Cluster analysis allowed the separation of six groups of homogeneous varieties with respect to their adaptability (the groups of varieties within the ellipses). Other varieties form single isolated clusters, and these are located on the edge of the six groups.

**References**


