# From Curves to Complexity: Functional and Topological Insights into Glycemic Control in Type 1 Diabetes

*Andrea Corradini*

The source of the data is Jaeb Center for Health Research (2017). The datasets have been retrieved from http://https://public.jaeb.org/dataset/546. The analyses content and conclusions presented herein are solely the responsibility of the authors and have not been reviewed or approved by the Jaeb Center for Health Research.

---

**Abstract**

In this study, we applied Functional Data Analysis and Topological Data Analysis techniques to continuous glucose monitoring (CGM) data from 225 adults with type 1 diabetes. We worked on a standardized set of daily blood glucose curves (288 points per curve), associated with clinical variables such as HbA1c, frequency of insulin boluses, and level of fear of hypoglycemia. Functional analysis included Functional Principal Component Analysis, clustering using k-means, functional boxplots (MBD and BD2), regression models (scalar-on-function and function-on-scalar) and FANOVA. Clustering identified four groups with distinct mean glycemic profiles and differentiated HbA1c values. Functional regression showed HbA1c to be the main predictor of glycemic trends. Topological analysis revealed simple structures in most groups, with greater complexities in subjects with suboptimal glycemic control.

---

## Introduction

In the past decade, data from continuous glucose monitoring (CGM) devices have become available. This has made it possible to analyze new perspectives in the study of type 1 diabetes. The data in question, which record blood glucose concentration (mg/dL) at regular intervals throughout the day, provide a way to visualize the trend of blood glucose curves with a very high degree of accuracy. The idea of this work is to apply Functional Data Analysis and Topological Data Analysis techniques to study the glycemic curves of a sample of adult patients with type 1 diabetes. The main objective is to navigate deeply into the functional characteristics of the curves and their topological structure to understand how they associate with relevant clinical variables, in particular the HbA1c value (i.e., glycated hemoglobin, which corresponds to the average glycemic value recorded over the past three months), the frequency of insulin bolus administration, and the level of fear of hypoglycemic events (blood glucose < 70mg/dL). The dataset analyzed is from the Awesome-CGM

public repository (Aleppo, 2017) and contains data from 225 adult patients with type 1 diabetes, with continuous glycemic records over 6 months and related clinical and behavioral information. After a pre-processing and cleaning phase, we selected a set of valid glycemic curves on a daily basis, interpolated on a uniform temporal grid with five-minute resolution (288 points per curve), in order to construct a consistent functional representation of the measurements.

What we are asking is: can daily patterns of glycemic curves in patients with type 1 diabetes reveal functional or topological signatures that can discriminate different levels of glycemic control, overcoming the limitations of traditional indicators such as HbA1c, and showing how behavioral variables interact with glycemic dynamics in determining the risk of poor control?

**Data Description**

As mentioned, the study is based on the analysis of data from a publicly available dataset described in the work of Aleppo et al. (2017). The dataset was originally collected in the context of the REPLACE-BG clinical trial, a randomized trial conducted in order to study what the advantages of self-management of blood glucose using CGM over traditional glucometer monitoring might be. Main files containing CGM data, insulin boluses (HDeviceBolus.txt), HbA1c values (HLocalHbA1c.txt) and questionnaire scores (HQuestHypoFear.txt) were uploaded for analysis. The CGM data record glycemic values at regular intervals, with timestamps that were converted to a full temporal format (POSIX datetime), reconstructing for each observation the exact time of day at which the measurement was taken. Records with no glycemic value were excluded to ensure the quality of the dataset. In the next step, a selection of data was made based on criteria of completeness and continuity. In order to reduce the impact of days with too piecemeal measurements, only days when at least 260 glycemic measurements were available were considered valid, corresponding to about 90% of the possible observations in a typical day with CGM at 5-minute intervals. In addition, to reduce the risk associated with an uneven presence of data among participants, only patients with at least seven valid days were included in the analysis.

The curves that passed the controls were then subjected to a process of imputation of missing data by linear interpolation and subsequent smoothing based on B-spline bases, with the number of bases chosen according to the temporal resolution and regularization required to limit overfitting. The final dataset thus represents a set of standardized daily glycemic curves, to which information on HbA1c value, frequency of insulin boluses administered, and average level of fear of hypoglycemia were associated for each patient.

**Methodology**

The analysis was structured in several steps, making use of advanced Functional Data Analysis and Topological Data Analysis techniques. First, the daily blood glucose curves were subjected to Functional Principal Component Analysis, with the aim of identifying the main patterns of variation of the curves within the sample. The FPCA method using the pca.fd function was applied, with function centering and retention of the first ten harmonics, which were able to explain a significant proportion of the overall variability of the curves. The trend of explained variance was represented by a scree plot, while the distribution of scores on the first two principal components was explored with a colored biplot as a function of HbA1c value.

Next, the scores of the first three principal components were used to perform clustering analysis by the k-means method. The optimal number of clusters was jointly determined through the elbow method (WSS plot) and silhouette analysis. Once the appropriate number of clusters was identified, the averages of the blood glucose curves for each group were estimated. A functional boxplot was also made, using both the Modified Band Depth (MBD) and BD2 methods, to provide a representation of the internal variability of the curves and identify any functional outliers. To further investigate the differences between groups, depth measures (MBD) were calculated and nonparametric statistical tests such as Wilcoxon were performed in order to compare depth values between groups of patients divided according to HbA1c level.

The methodology also involved the application of regression models to analyze the relationships between the characteristics of blood glucose curves and clinical variables. Linear models were estimated that considered the scores of the first three principal components as dependent variables and the standardized values of HbA1c, the number of daily insulin boluses and the average level of fear of hypoglycemia as explanatory covariates. In parallel, functional regression models were estimated. The first scalar-on-function model considered the standardized value of HbA1c as the response variable and the entire glycemic curve as a functional predictor. A second model extended the analysis by including scalar covariates (number of boluses and fear of hypoglycemia), modeled as constant predictors over time. Finally, a function-on-scalar model was applied in which the glycemic curve was assumed as the response variable and the scalar predictors were modeled as effects on B-spline functions along the time domain. To test for significant differences in the time domain between groups, two functional analysis of variance (FANOVA) analyses were performed. The first compared curves between clusters identified by clustering, while the second compared curves between patients with different HbA1c. In both analyses, the trend of the F-statistic along the time domain was calculated and the time slots where the differences were most significant were identified.

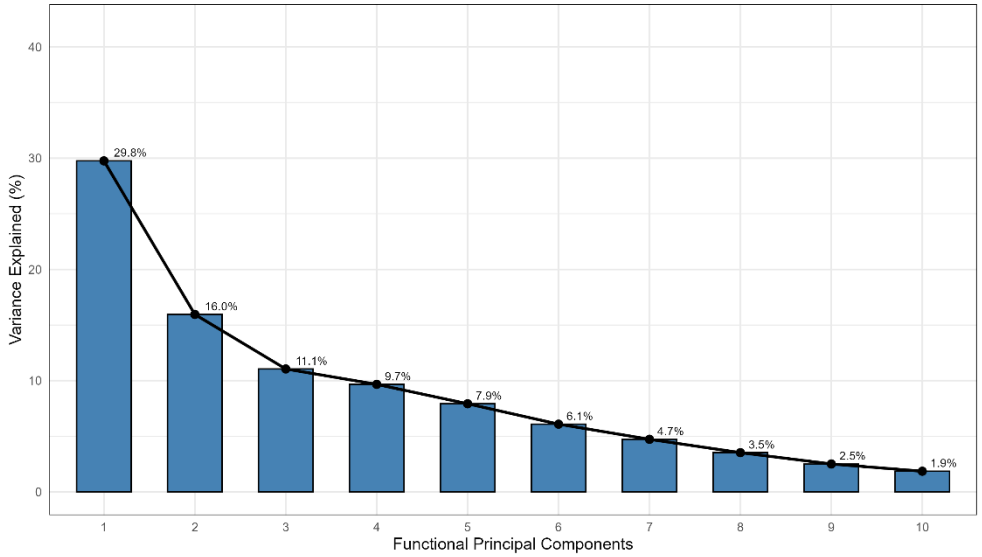
As part of the Topological Data Analysis, persistence plots were constructed for the blood glucose curves belonging to the groups of patients with HbA1c differentiated into four different categories. The analysis was conducted based on the construction of alpha complexes and identified

the topological features associated with each group. Bottleneck distance in $H_0$ dimension was calculated to quantify the topological difference between pairs of curves and provide a summary measure of dissimilarity in their connective structure.

## Results

The main results from the different stages of the analysis are presented below, accompanied by graphical representations to facilitate their interpretation. Figure 2 shows the scree plot related to the functional principal component analysis.

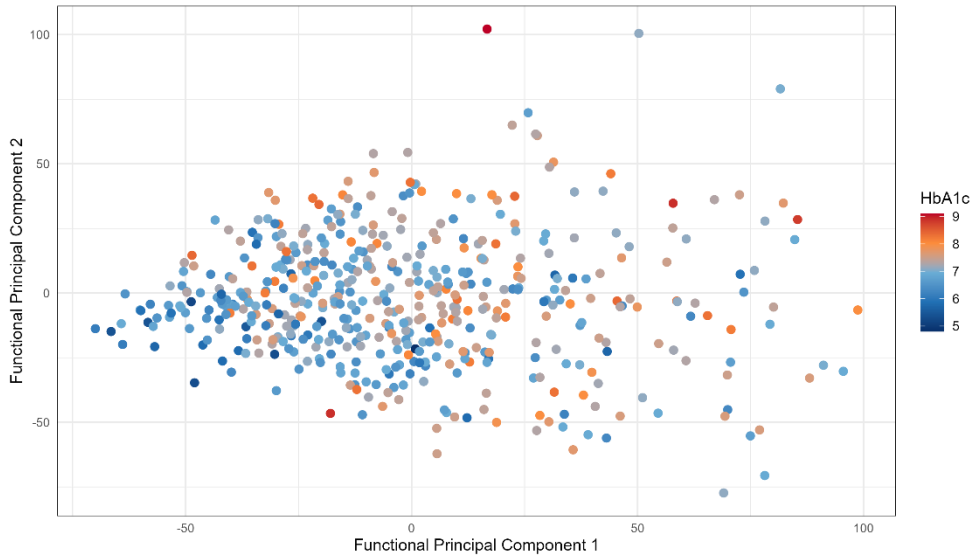
Fig. 2 – Scree Plot




The first three components together explain 57 percent of the total variance of the daily glycemic curves: the first principal component captures 29.9 percent of the variability, the second 16 percent, and the third 11.1 percent. Subsequent components contribute progressively lower percentages of variance. Except for Figure 4, where we make use of the fact that the first five components explain more than 70 percent of the variance to visualize the first five harmonic functions obtained from FPCA, the above confirms the goodness of the choice to focus the analysis on the first three principal components. This level of variance explanation represents a compromise between synthesis and descriptive ability, consistent with what has been observed in the literature for glycemic data.

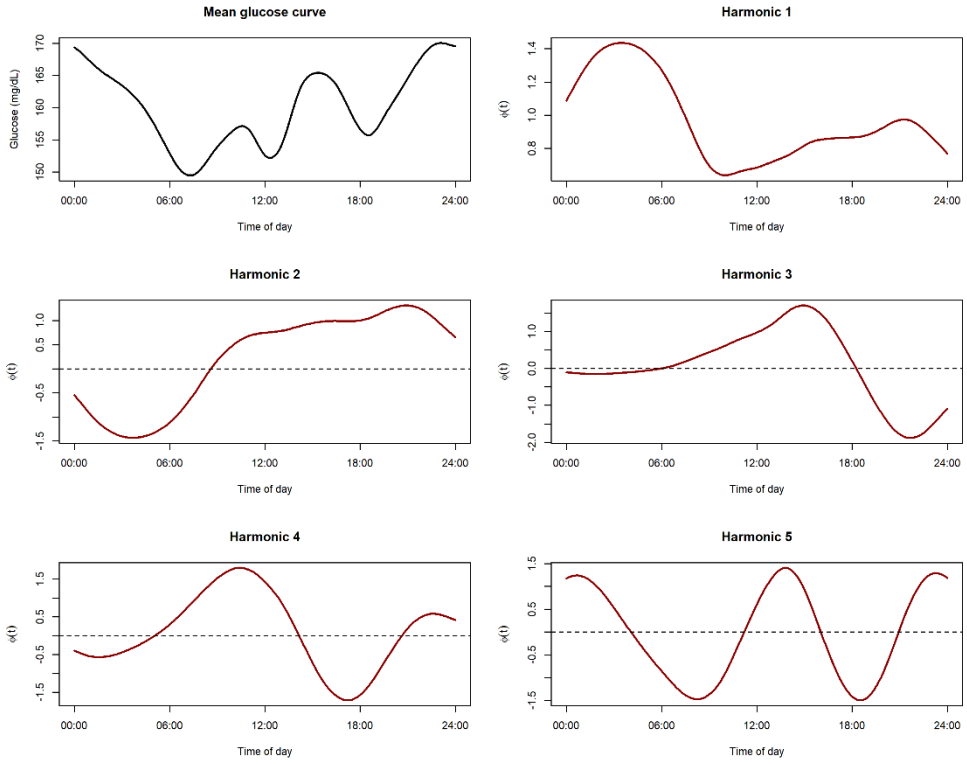
Figure 3 shows the distribution of the first two principal component scores obtained from FPCA on a sample of 500 curves, where each point represents a daily glycemic curve. The horizontal axis corresponds to the score on the first principal component (PC1), while the vertical axis represents the score on the second principal component (PC2). The points are colored according to the associated HbA1c value, according to a continuous scale ranging from the lowest (dark blue) to the highest (red) values.

Fig. 3 – FPCA Scores (PC1 vs PC2 colored by HbA1c)




It is apparent from the graph and the data that the scores are distributed over a relatively wide space, with PC1 ranging approximately between -70 and +100 and PC2 between -70 and +100. HbA1c values in the sample are predominantly in the range of 6.0-8.0, with some cases above 8.0 showing up as red dots. Some heterogeneity is observed: the curves associated with higher HbA1c do not cluster in a specific area of the plane, but are distributed over different combinations of PC1 and PC2.

Figure 4 – Harmonics plotted




This suggests that the variability of glycemic forms captured by the first two principal components is not strictly dominated by HbA1c. The staining shows some spread of HbA1c values over the principal

components, emphasizing that HbA1c reflects a long-term average that may not be perfectly discriminated by the first two functional components alone.

Figure 4 presents the average glycemic curve and, as anticipated, the first five harmonic functions obtained from FPCA. The average curve depicts a typical glycemic profile over the 24-hour period: a morning minimum is observed, indicative of stability at night and in the early hours of the day, followed by two main peaks-the first in the middle hours (probably related to the mid-day meal) and the second in the evening range, near dinner. The first harmonic represents the largest component of variability. It shows a marked contrast between morning and evening values: positive or negative values on this component indicate curves with generally higher or lower-than-average blood sugar levels at these times of day. Harmonic 2 and Harmonic 3 capture differences between glycemic behavior during the middle hours of the day and the extremes of the day. Harmonic 4 and Harmonic 5 are indicative of subjects with glycemic curves characterized by sudden fluctuations or greater instability throughout the day.

Following the functional analysis and dimensional reduction by FPCA, a clustering analysis was conducted by the k-means method, using the scores of the first three principal components as input. The analysis was performed on a sample consisting of 500 daily blood glucose curves. The choice of the optimal number of clusters was supported by multiple diagnostic tools. The plot of WSS, or the sum of intra-cluster distances, shown in Figure 5, shows how the increase in the number of clusters is associated with a progressive reduction in WSS. The WSS curve shows a clear inflection point at the fourth cluster, suggesting that adding more clusters would result in a marginal benefit in terms of reducing intra-cluster variability.

Fig. 5 – WSS Plot (Elbow Method)

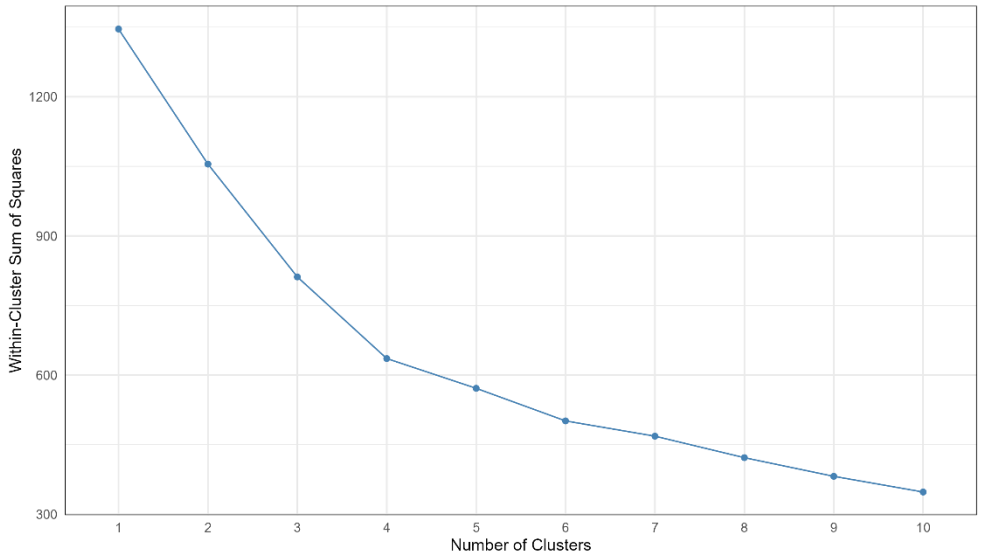
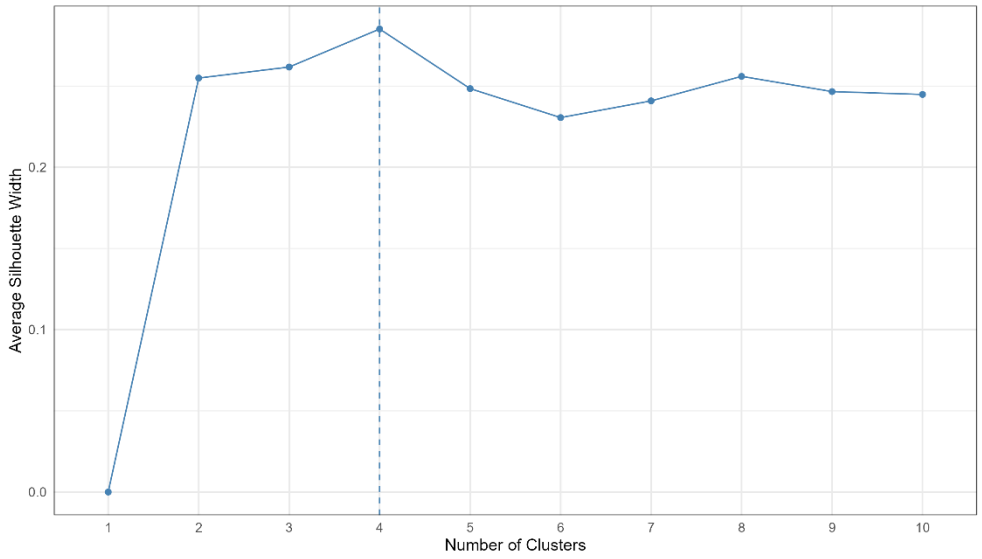


Fig. 6 – Average Silhouette Width




The evaluation was further strengthened by silhouette analysis, depicted in Figure 6 and Figure 7. Figure 6 shows the average value of the silhouette coefficient for each number of clusters considered. The maximum value is observed with four clusters, although the index does not exceed 0.3. Figure 7, which shows the distribution of silhouette coefficients for each cluster, allows us to appreciate how cluster 3 exhibits greater internal cohesion than the other clusters, with an average silhouette of 0.38. The remaining clusters show lower average values, hovering around 0.18-0.20, suggesting greater internal heterogeneity.

Fig.7 – Silhouette Width

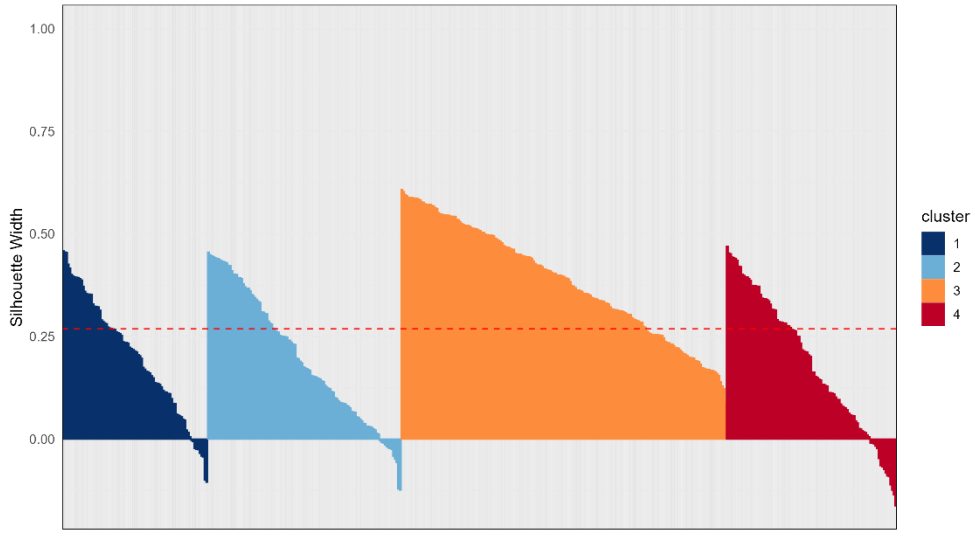



Figure 8 proposes a two-dimensional representation of the data in the space of the first two principal components, with the points colored according to the cluster they belong to. Observation of the graph confirms what emerged from the quantitative indices: some clusters show greater spatial compactness, as in the case of cluster 3, while others are more diffuse or overlapping, particularly in

the central regions of principal component space. This distribution reflects the complex nature of glycemic curves, in which distinctive patterns are not always markedly separable in a two-dimensional plane.

Fig. 8 – K-Means scatterplot

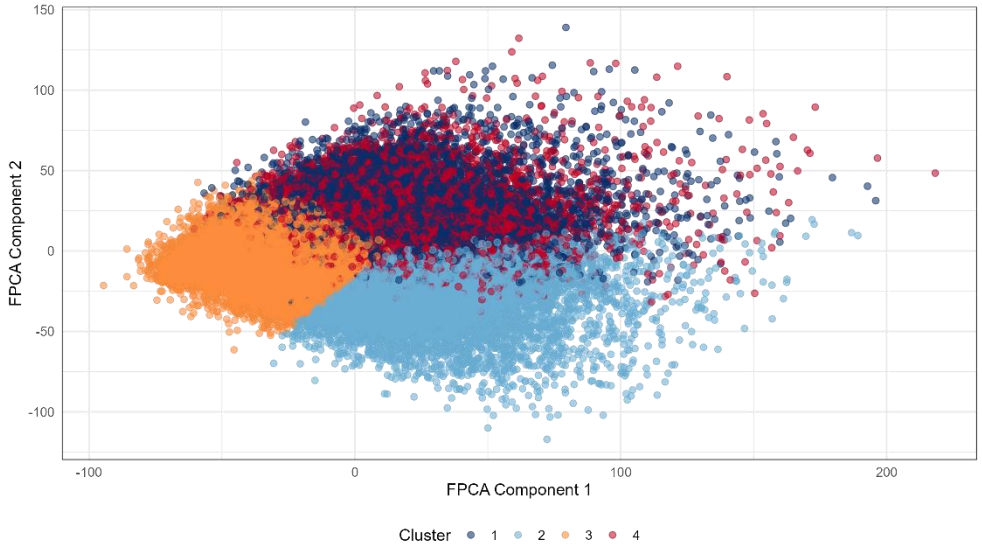
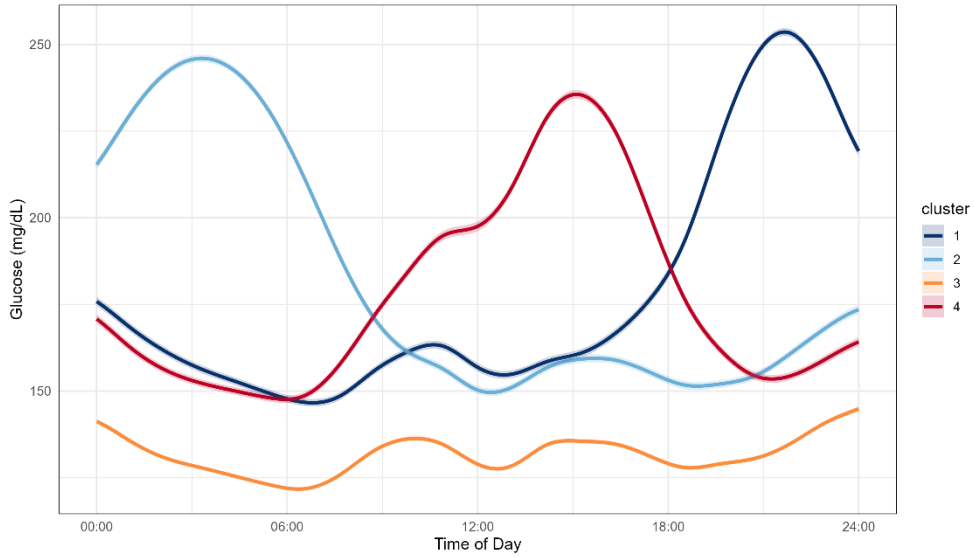


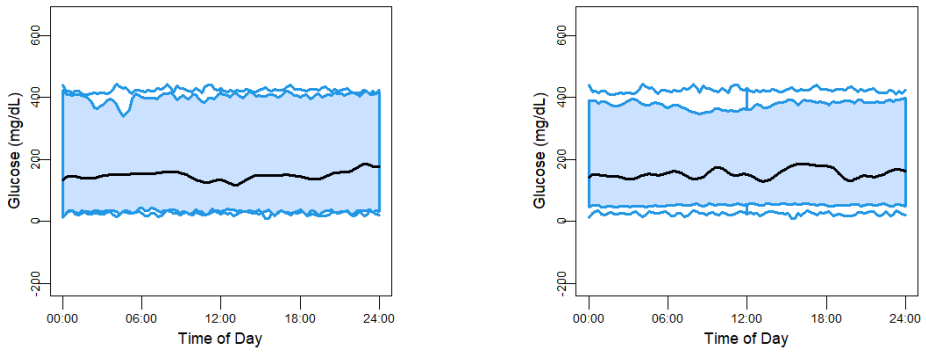

Figure 9 shows the averages of the blood glucose curves for each cluster, accompanied by the 95% confidence intervals. Distinctive mean profiles characterizing each cluster are observed. Cluster 1 shows a glycemic pattern with relatively high average values at night and marked peaks in the middle and evening hours of the day. Cluster 2 is distinguished by flatter average curves with lower glycemic values throughout the day. Cluster 3, which as anticipated represents the most cohesive group according to the silhouette analysis, shows a regular profile with well-defined peaks at main meals. Finally, cluster 4 shows a similar pattern to cluster 1, but with slight variations in postprandial peaks and greater variability at night.

Fig. 9 – Cluster-wise Mean Glucose Curves



Cluster 3 not only collects the most curves, but is also the one associated with the lowest mean HbA1c value (6.81), consistent with better long-term blood glucose control. Clusters 1 and 4 have substantially equivalent mean HbA1c values (7.26 and 7.25 respectively), while cluster 2 shows a slightly lower mean HbA1c (7.17).

Figure 10 – Functional Boxplots (MBD vs. BD2)




In our study, we used functional boxplots as an exploratory tool to describe the distribution of daily blood glucose curves over the entire data set. The boxplots, constructed using two different depth measures, Modified Band Depth (MBD) and second-order Band Depth (BD2), allowed us to graphically represent the functional centrality of the curves and their variability throughout the day. In both boxplots in Figure 10, the black line highlights the functional median, while the blue band identifies the central box enclosing the most representative curves. In both cases, it can be seen that the central curves are relatively stable throughout the day, while the outliers are clearly visible outside the limits of the box. It can be seen that the MBD tends to favor a more robust identification of the global centrality of the glycemic profile, while the BD2 appears more sensitive to local variations in the curves.

For a more complete and representative analysis of the depth distribution, Modified Band Depth (MBD) values were calculated on the entire set of available curves. The histogram shown in Figure 11 shows that most of the curves have intermediate depth values, with an obvious peak around 0.7, while there is no shortage of curves characterized by lower MBD values. This trend confirms the coexistence of a compact core of central curves and a set of profiles that are more atypical than average.

Fig. 11 – MBD Values Histogram

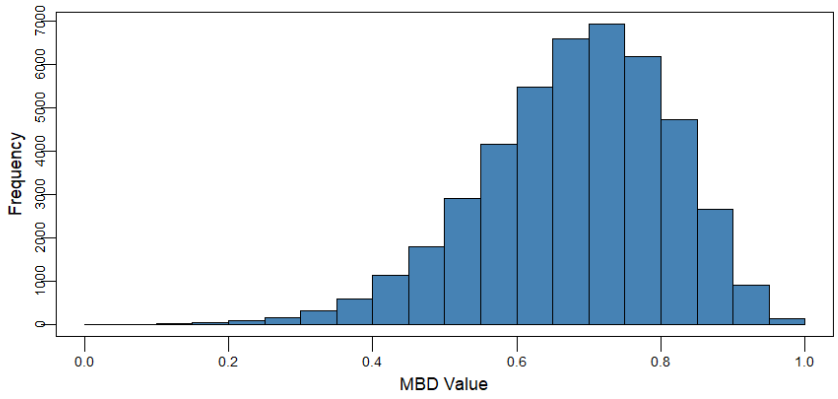
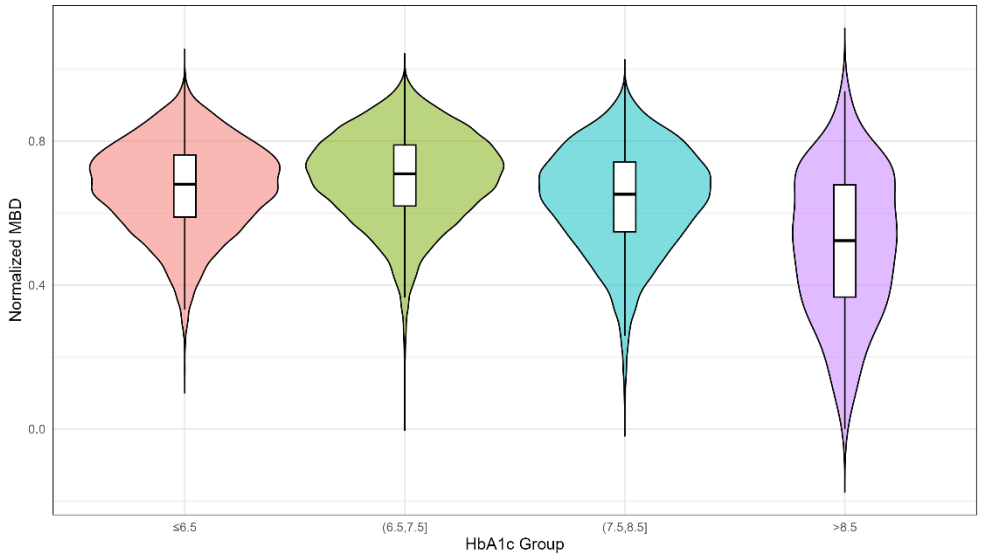


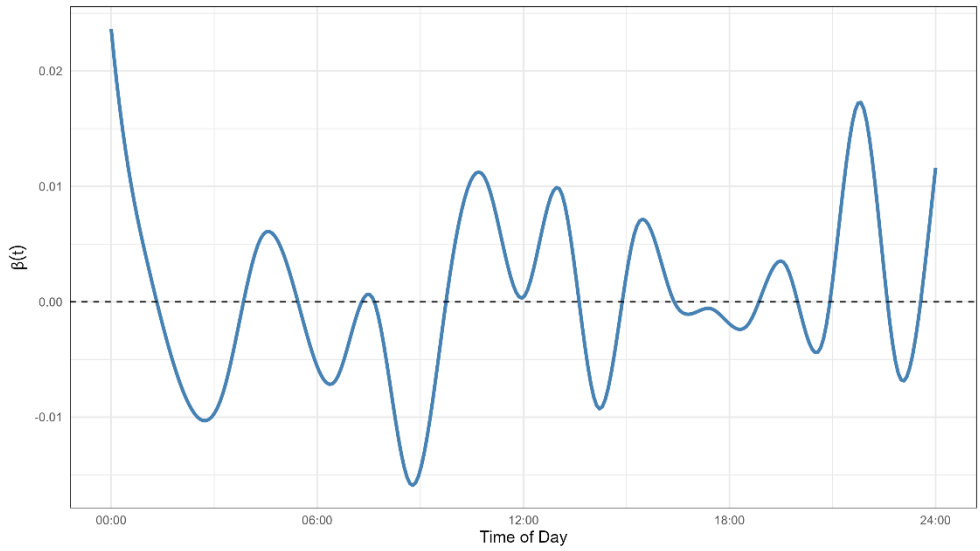

Fig.12 – MBD by HbA1c Group




In order to explore the relationship between MBD and glycemic control, the analysis was conducted by considering the entire set of curves and dividing the data into four categories based on HbA1c values: subjects with values ≤6.5, subjects with values between 6.5 and 7.5, between 7.5 and 8.5, and subjects with values greater than 8.5. The violin graph (Figure 12) visually depicts the distribution of MBD values within each group and allows one to immediately grasp the differences between the categories. The shape and width of the violins clearly show how subjects with lower HbA1c tend to have higher and more concentrated MBD values. In contrast, as HbA1c values increase, there is a progressive widening of the distribution and a shift toward lower MBD values. This trend manifests itself clearly and uniformly across the scale of HbA1c categories, visually confirming the association between glycemic control and functional centrality of the curves.

These graphical observations were supported by statistical analysis: the Kruskal-Wallis test returned a p-value < 2.2e-16, showing globally significant differences between the groups. Examination of pairwise comparisons by Wilcoxon rank-sum test, with correction for multiple comparisons according to the Benjamini-Hochberg method, further confirmed these differences, showing p-values less than 2e-16 for all combinations of group pairs.

Fig. 13 – Functional Coefficient (HbA1c ~ curve only)




The functional analysis begins by estimating the functional coefficient related to HbA1c alone, depicted in Figure 13. The function shows noticeable fluctuations at different times, with a greater influence in the early morning and evening periods, suggesting that the mean glycate is associated with an increase or decrease in the glycemic curve at specific times, probably related to the typical eating and insulin rhythms of the day. The simple linear model shows a significant and robust relationship between HbA1c and the first principal component of the glycemic curves, with a coefficient of about 20 and an $R^2$ around 15 percent, indicating a discrete ability of glycation to explain the overall variability of the curves.

When the standardized covariates n_boli_z and mean_fear_z are added to the model, as shown in Figure 14, the functional coefficient of HbA1c changes slightly. It can be seen that the addition of these variables allows some of the variability previously attributed to glycation alone to be absorbed. The estimated function appears less pronounced in some traits, a sign that the frequency of insulin boluses and the average level of fear help explain some of the link between glycate and the glycemic curve. The contribution of glycate remains dominant, but there is evidence of a significant role of the other two covariates as well, especially for the first principal component. On the second component, glycate continues to have a significant, albeit small, effect, while n_boli_z shows a significant but more modest association, and mean_fear_z appears less impactful. On the third component, finally, the estimated effects are weaker and less relevant from a practical point of view, as shown by the value of $R^2$ close to zero. The combination of functional approaches and regression on principal components thus allows for an articulated reading of the link between clinical characteristics and blood glucose trends over time.

Fig. 14 – Functional Coefficient (HbA1c ~ curve + covariates)

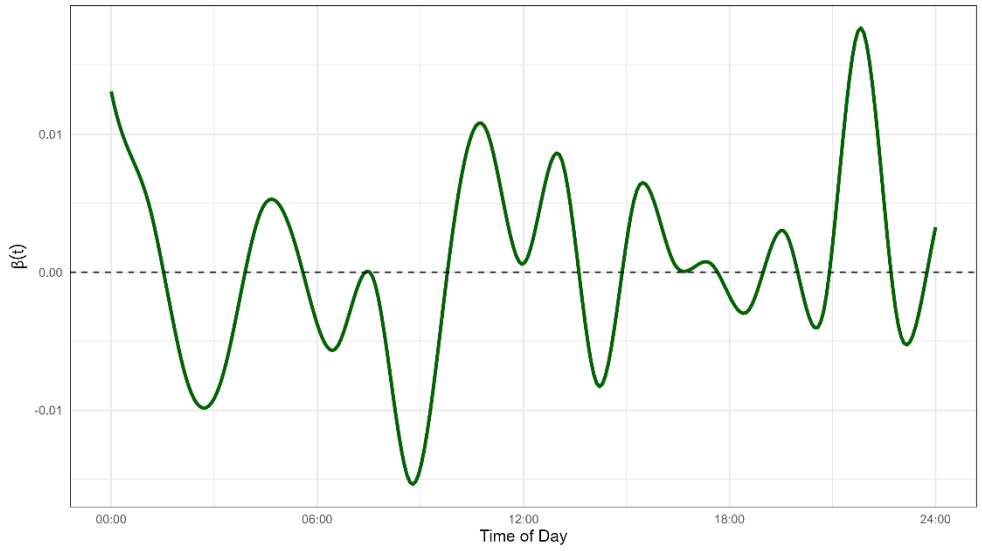
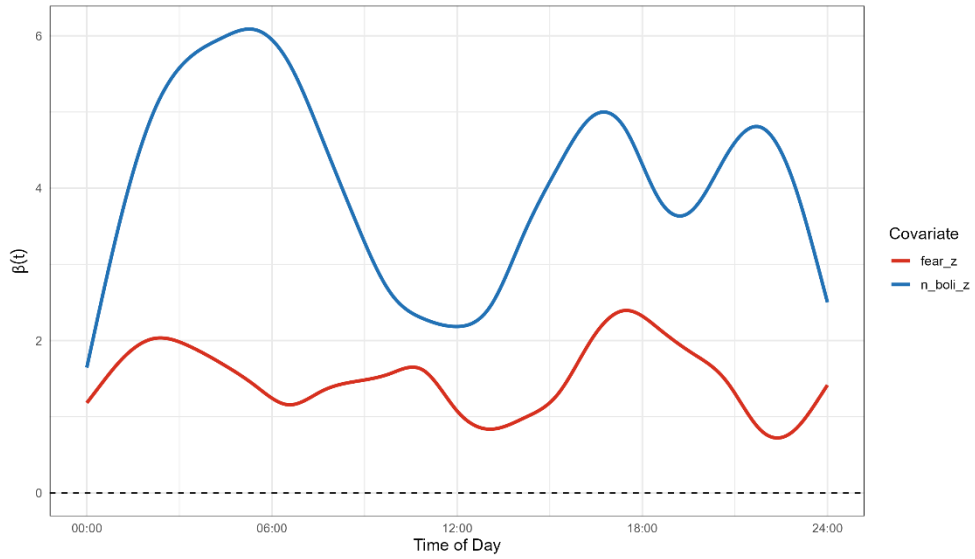



Figure 15, related to the function-on-scalar model, summarizes the estimated effects of n_boli_z and mean_fear_z over time, visually confirming the dynamics described. The effect of n_boli_z exhibits peaks at times when meal intake and insulin administration are most likely, while the effect of fear remains at smaller but still nonnegligible values at some times of the day.

Fig. 15 – Function-on-scalar: effects over time

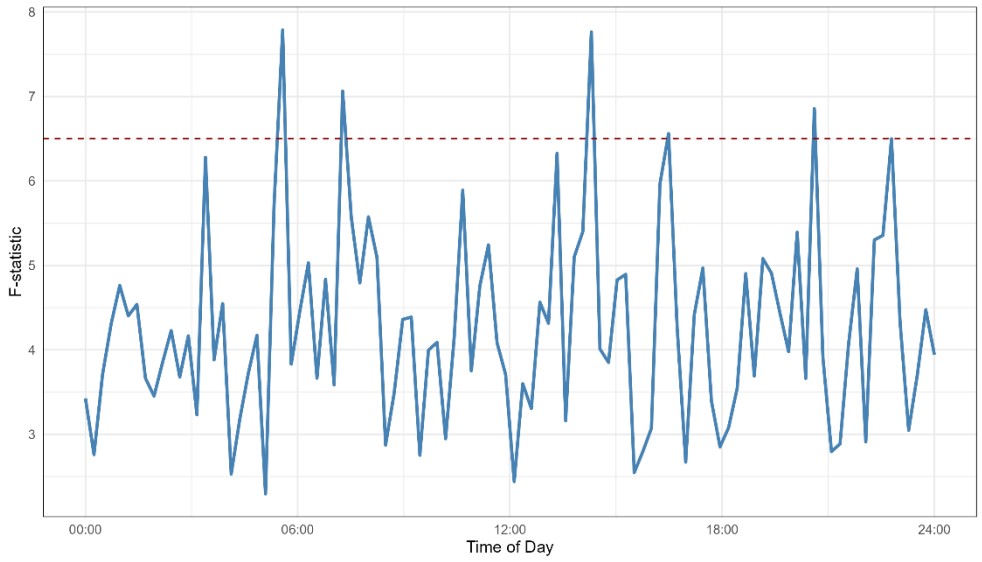



The results of the cross-validation conducted on a random subsample of 1,000 blood glucose curves provide interesting insights into the predictive ability of the models tested, although they highlight some critical issues. The first model returned a cross-validation $R^2$ value of approximately -0.0087. The negative value of $R^2$ indicates that the sum of squares of errors predicted by the model exceeds that of a model without predictors. The second model, which enriches the functional regression by including two scalar covariates, showed an $R^2$ from cross-validation of about -0.0104. Again, the result shows that the addition of the scalar variables does not lead to an improvement in

the overall predictive ability of the model on the subsample used. Rather, the slight worsening of $R^2$ could reflect excess complexity that is not justified in the data, or it could depend on noise or weak relationships between predictors and response that do not generalize well to the validation data. Overall, the cross-validation results suggest that the models tested, in the configurations adopted, fail to capture sufficiently robust relationships between daily glycemic trends and HbA1c levels from a predictive point of view.

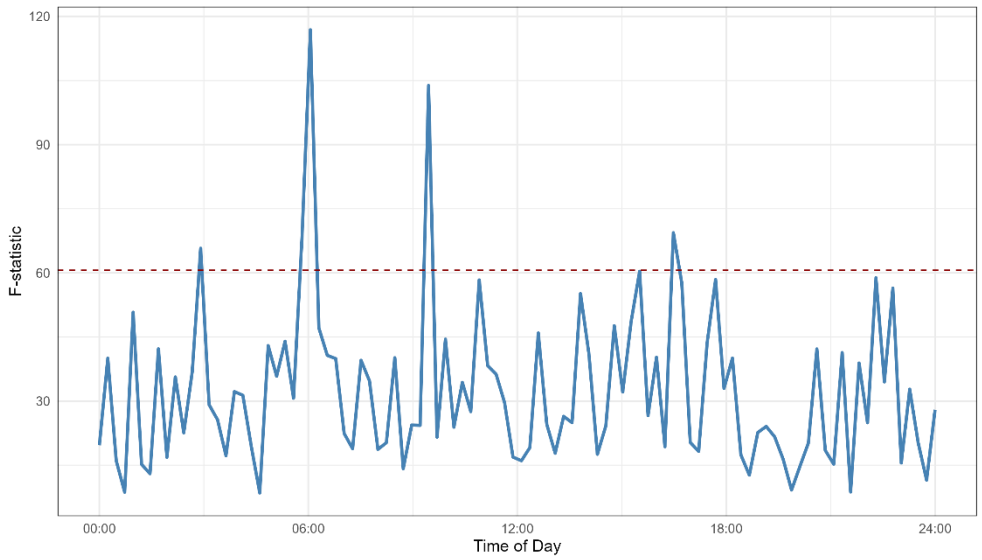
Let us now turn to a discursive commentary on the results obtained through the application of FANOVA, with reference to the two graphs produced. In the first graph (Figure 16), concerning the comparison of the glycemic curves divided by clusters identified by means of K-means, the F-statistic remains below the critical threshold for most of the day, indicating no statistically significant differences over most of the time interval. However, a few rare peaks exceeding the threshold are observed, located mainly in the postprandial and nocturnal periods of the day. These peaks suggest that weak signs of differentiation between clusters might emerge at short times of the day, but their sparseness and patchy distribution mean that these differences are not systematic or robust enough to support a robust functional interpretation.

Fig. 16 – FANOVA (F-Statistic by Cluster)



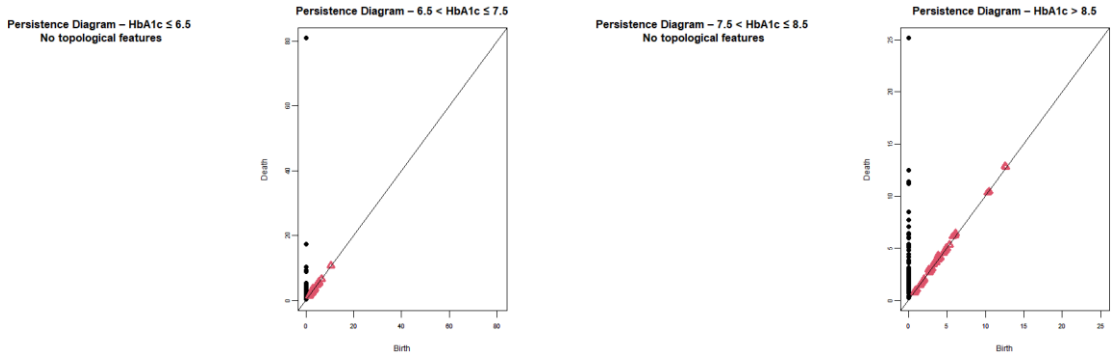

In Figure 17, which analyzes the differences between the glycated hemoglobin groups divided into four categories, the trend of the curve shows numerous fluctuations and some peaks that reach or exceed the critical threshold indicated by the dotted line. These peaks are mainly concentrated in the early morning hours and some time slots in the afternoon, going to highlight dynamics probably related to meals and/or insulin management. In spite of this, the detected signals are nevertheless isolated and do not configure a uniform or systematic functional pattern, leaving open the hypothesis that part of the observed variability may be due to random fluctuations or local phenomena that cannot be easily generalized to the entire population analyzed.

Fig. 17 – FANOVA (F-Statistic by HbA1c Group)




Finally, we turn to Topological Data Analysis (TDA), an innovative approach that allows us to explore the inherent structure of data without imposing rigid models. TDA is based on the study of topological properties, such as connected components and cycles, that emerge when data are progressively "connected" by a filtering parameter. The goal, in our case, is to investigate whether glycemic curves show relevant structural differences between groups of subjects divided according to HbA1c values. The graph of persistence plots is presented in Figure 18: to make it, we selected a sample of 100 curves for each group (which will be repeated for the bottleneck distance). It is immediately observed that in the groups with HbA1c $\leq$ 6.5 and 7.5 < HbA1c $\leq$ 8.5, no persistent topological features were detected: the diagrams are either empty or lack significant points outside the diagonal. This indicates that the glycemic curves in these groups did not generate complex topological structures in terms of connected components that persisted as the filtering threshold changed. In contrast, in the 6.5 < HbA1c $\leq$ 7.5 and HbA1c > 8.5 groups, some connected components with greater persistence emerge, as shown by the points further away from the diagonal. These features signal the presence of more complex or fragmented local patterns in these groups.

Fig. 17 – Persistence diagrams



Quantitative comparison by bottleneck distance confirmed this observation: the only calculable distance was between the $6.5 < HbA1c \leq 7.5$ and $HbA1c > 8.5$ groups, with a value of about 40.49. A distance of this magnitude suggests the presence of significant structural differences in the distribution of the connected components between these two groups, probably related to greater complexity or variability of the curves in subjects with worse glycemic control. For the other pairs, the bottleneck distance was not calculated because at least one of the plots was devoid of persistent features, rendering the topological comparison meaningless. Overall, TDA provided a picture consistent with what emerged from the functional analyses: the blood glucose curves exhibit a simple topological structure in most groups, while some differences emerge only among specific HbA1c categories. These results indicate that topological complexity is not a dominant feature in the glycemic curves of the entire sample, but it may be more pronounced in subjects with suboptimal glycemic control.

**Conclusions**

Our work demonstrated how the integrated use of Functional Data Analysis and Topological Data Analysis enables an in-depth exploration of daily glycemic curves in adults with type 1 diabetes, revealing functional and topological patterns associated with different levels of glycemic control. Functional analyses revealed that groups of curves with distinct mean profiles show different average HbA1c levels, suggesting that patterns in the glycemic curves are associated with glycemic control, although individual predictive power remains limited. This highlights potential limitations of relying exclusively on static indicators like HbA1c. The inclusion of behavioral variables, such as the frequency of insulin boluses and the level of fear of hypoglycemia, provided further insights into how patient behaviors interact with glycemic dynamics to influence risk profiles. Topological analyses identified increased complexity in the curves of patients with poorer glycemic control, suggesting that topological features may serve as novel indicators of suboptimal management. Although predictive performance in cross-validation was limited, these results underscore the potential of functional and topological approaches to contribute to a more personalized and dynamic understanding of glycemic control in type 1 diabetes, paving the way for future tools that move beyond static metrics toward continuous, curve-based monitoring.