Functional Data Analysis in Econometrics

Kateřina FRONČKOVÁ* and Pavel PRAŽÁK

University of Hradec Králové, Hradec Králové, Czech Republic; katerina.fronckova@uhk.cz; pavel.prazak@uhk.cz

* Corresponding author: katerina.fronckova@uhk.cz

Abstract: Functional data analysis is a modern and dynamically developing field of statistics, which finds its application in many different areas. It can also be used in econometrics. The paper explains the main ideas of functional data analysis and presents its fundamental theoretical background. It also shows an overview of the possibilities of using a functional approach in the analysis of economic and financial data through a search and review of scientific articles. Out of the total number of 1,492 articles dealing with the topic of functional data analysis and meeting the search criteria, 55 articles focusing on the solution of econometric problems are identified. Subsequently, the use of functional data analysis in the exploratory analysis of the industrial production index of the Czech Republic is demonstrated. The so-called phase plane plots are used for the analysis, which make it possible to better examine the dynamics of the development of industrial production. The analysis studies the development of production in individual years and examines the impact of various economic shocks. Using a functional approach, it is possible to answer questions that are difficult to answer using classical statistical methods.

Keywords: functional data analysis; econometrics; phase plane plot; industrial production index

JEL Classification: C19

1. Introduction

Functional data analysis is a modern and dynamically developing field of statistics. It has been booming over the last twenty years, so it is a relatively new approach compared to classical statistical methods known for centuries.

The subject of interest of functional data analysis is data that can naturally be perceived as functions. Although these data are available in practice in the form of sequences of individual observed values, the basic idea of functional data analysis is to approach these sequences as entire entities – functions (it is usually assumed that functions are continuous and smooth, and their value exists for all points of a certain domain on which they are defined, but they are observed only at specific individual points). The function is then the basic unit of analysis. The most common type of functions are functions with a one-dimensional continuous argument, which is typically time, but the argument can also be multidimensional, such as position in space, etc.

Functional data analysis combines the ideas and approaches of classical statistics and mathematical functional analysis. The goals of functional analysis are similar to that of other

areas of statistics, for example, to discover, describe and explain various important data characteristics, examine relationships between variables, create statistical models, perform inference, etc. The advantage of functional data analysis over classical statistics is the more natural perception of functional data, which makes it possible to use more information in analyzes concerning the behavior of a function or work with derivatives of a function. Moreover, functional data can be problematic from the point of view of classical statistics.

Functional data analysis finds its application in various natural and social sciences and applied domains. Its importance is closely related to the development of computer technology, which allows the acquisition, storage, and processing of large amounts of data of a functional nature.

The paper deals with the presentation of the fundamental ideas of functional data analysis and discussion of the possibilities of its use in econometrics. Section 2 is devoted to a review of scientific articles focused on the functional analysis of economic and financial data. Section 3 provides an overview of the main theoretical basis of functional data analysis. In section 4, the functional approach is used to analyze the dynamics of the industrial production index. Section 5 is the concluding summary.

2. Review of the Possibilities of Using Functional Data Analysis in Econometrics

To explore the possibilities of using functional data analysis in econometrics, a review of scientific articles was performed.

Articles were searched in the Web of Science and Scopus databases using the query string "functional data analysis", the occurrence of which was searched in article titles, abstracts, and keywords. Only articles from peer-reviewed journals and written in the English language were selected. The search was conducted in December 2020, so all articles published (and indexed in the database) up to that time were included. A total of 2,929 articles matching the search criteria were found in both databases. Of these, 1,437 records were duplicates. After their removal, 1,492 articles remained, the focus of which was studied.

The result was the identification of 55 articles describing the practical application of functional data analysis in the field of economics. The articles came from the years 2002–2020, in recent years the number of articles has been higher. The usual type of problem solved was an exploratory analysis, regression analysis, or time series analysis; in the tasks of creating various models, besides describing the relationships between variables, the most common goal was to predict future behavior.

The articles focused on the description and modeling of the behavior of various economic or financial variables and the examination of the influence of various factors on their behavior. Specifically, the subject of analyzes was, for example, the behavior of exchange rates (Kearney et al., 2018, 2019; Kim & Jung, 2018b), interest rates (Caldeira et al., 2020), price formation in online auctions and e-commerce (Bapna et al., 2008; Oomen, 2019; Reddy & Dass, 2006; Reithinger et al., 2008; S. Wang et al., 2008), prices of strategic commodities such as oil (Kearney et al., 2015; Kearney & Shang, 2020; Kim & Jung, 2018a), real estate prices (Peng et al., 2014; Seya et al., 2016), wages and wage inequality (Scott & Handcock, 2005; D. Wang et al., 2018), behavior of stock indices (Kokoszka et al., 2017; Müller et al., 2011; Z. Wang et al.,

2014), modeling and forecasting stock or bonds returns (Brockhaus et al., 2018; Caldeira & Torrent, 2017; Cao et al., 2020; Das et al., 2019; Hays et al., 2012), including the related quantification of investment risk and portfolio optimization (Cai, 2018; Kokoszka et al., 2019). The dynamics of supply and demand in various markets was also examined (Canale & Vantini, 2016).

3. Theoretical Fundamentals of Functional Data Analysis

Some basic theoretical aspects of functional data analysis are presented, more detailed information can be found, for example, in (Hsing & Eubank, 2015; Kokoszka & Reimherr, 2017; Ramsay & Silverman, 2005).

3.1. Functional Random Variable

A functional random variable can be formally defined as a variable that takes values in an infinite dimensional space (functional space), $X = \{X(t) : t \in T\}$. In functional analyzes, we work with observations of the functional variable resp. functional data, $\{x_i(t) : t \in T, i = 1, 2, \ldots, N\}$. The function x_i is the basic object (unit) of analyzes, it is assumed that its value $x_i(t)$ exists for all $t \in T$, but in practice it is observed only in specific discrete points $t_{ij} \in T, j = 1, 2, \ldots, n$. The argument t is often one-dimensional (it is then a function of one variable, typically for example time), but it can also be two-dimensional or generally multidimensional. Further in the text, it will be assumed that $t \in T \subset \mathbb{R}$.

3.2. Creating a Functional Representation of Data

Since functional data are available in practice only in the form of individual discrete observations, it is first necessary to create their functional representation. If it is assumed that these observations are not loaded with noise, it is an interpolation problem. On the other hand, if noise, such as measurement errors or various random perturbations, is present in the observations, it is a smoothing problem. In general, the following model can be considered

$$y_{ij} = x_i(t_{ij}) + \epsilon_{ij}, \tag{1}$$

where $x_i(t_{ij})$ is the actual value of the function x_i at the point t_{ij}, y_{ij} is the corresponding observed value and ϵ_{ij} denotes possible (additive) noise.

The most common way of representing a function is to express it in the form of a linear combination of certain basis functions

$$x_i(t) \approx \sum_{k=1}^{m} c_{ik}\phi_k(t), \tag{2}$$

where ϕ_k is a certain standard collection of basis functions (e.g., splines, sine and cosine functions, wavelets, etc.) and c_{ik} are the corresponding coefficients of the linear combination.

The representation is chosen so that the function x_i suitably corresponds to the observed data, is continuous and usually also smooth; everything depends on the requirements of the practical problem that is solved.

3.3. *Characteristics of Functional Data, Derivatives of Functions, Functional Models*

As in classical statistics, it is possible to define basic descriptive characteristics (such as mean, variance, covariance, correlation) also for functional data, with the only difference that they have the form of functions.

As already mentioned, the usual requirement is the smoothness of functions, which is also associated with the existence of their derivatives. Depending on the required number of derivatives, the basis for representing the functions, etc. is then chosen.

The functional approach can be useful in exploratory analysis, but functional variables can also appear in various models, purely functional or in combination with classical scalar variables, and the derivatives of functions can also appear in the models.

4. Demonstration of the Application of the Functional Approach in the Analysis of the Industrial Production Index

The use of functional data analysis in economics will be shown on the exploratory analysis of the industrial production index. A functional representation of the data will be created, and the derivatives of the function will be used in the subsequent analysis.

4.1. *Industrial Production Index and Data Source*

The industrial production index measures the output of the industrial sector adjusted for price effects. It is a basic indicator of industrial short-term statistics. Its calculation is given according to the Statistical classification of economic activities in the European Community, (NACE Rev. 2, Eurostat). Data on the industrial production index of the Czech Republic in the years 2000–2020, which was obtained from the website of the Czech Statistical Office, were used for the analyzes. The data took the form of monthly indices calculated against the base index, which is the average index of the year 2015 (index 2015 = 100). A total of 252 values were involved, the values were adjusted for the effect of calendar variations, the logarithms of the indices were used in subsequent analyzes. The whole time series can be seen in Figure 1, a growing trend and seasonal variation are evident. The implementation of analyzes was performed in R.
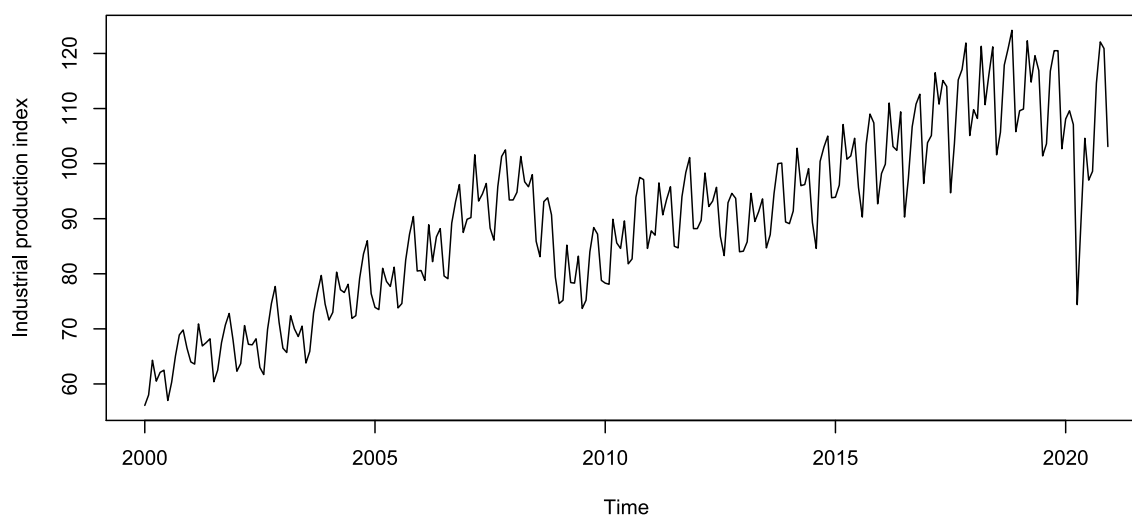

Figure 1. Industrial production index of the Czech Republic in the years 2000–2020.

4.2. Functional Representation

Industrial production can be seen as a continuous-time process, even if the index is given only at discrete points. To create a functional representation of the data, a B-spline basis was chosen, which is suitable for this type of data. The splines were defined by knots that were placed in each data point. For the construction of splines, polynomials of order six (degree five) were used, which provide sufficient flexibility to represent the behavior of the function and at the same time allow to obtain a continuous estimate of the first four derivatives of the function.

Roughness penalty smoothing was used to calculate the coefficients of the linear combination (2). This approach is based on the least squares method, but the aim is not only to minimize the sum of squared errors (SSE) but to minimize the characteristic

$$PENSSE = \sum_{j=1}^{252} [y_j - x(t_j)]^2 + \lambda \int_{2000}^{2020} [D^4 x(t)]^2 \, dt,$$

where $\lambda \int_{2000}^{2020} [D^4 x(t)]^2 \, dt$ is a penalty term that penalizes an overly complex and variable (rough) behavior of a function and ensures that the estimated function is naturally smooth (in general, this approach allows to capture only the essential characteristics of the behavior of the function without the influence of random perturbations and noise in the data). The fourth derivative of the function $D^4 x(t)$ was used for the penalty; the second derivative of the function represents its curvature and is usually chosen as a measure of the roughness of the function, but in subsequent analyzes the first and second derivatives of the function will be used, and it is required that also their behavior is not too rough, therefore, the second derivative of the second derivative of the function i.e. the fourth derivative of the function is used for the penalty (for this reason, polynomials of order six were used for spline construction). The smoothness or roughness of the function is now given by the choice of the smoothing parameter λ. The higher the value of λ, the stronger the effect of the penalty, and the resulting estimate of the function will be smoother, and vice versa, the smaller λ (closer to zero), the greater complexity of the estimated function will be allowed. In practice, the parameter λ is often chosen automatically using cross-validation, but here its value was determined manually so that the smoothness of functions and their derivatives is appropriate to the needs of subsequent analyzes (the value obtained by cross-validation would over-smooth the estimates).

The obtained estimate of the function and both derivatives is shown in Figure 2 for the year 2015 as an example.

The graph of the function shows the seasonal character of the industrial production index, the index reaches higher values in March, June, and October, on the contrary, the decline in production occurs in the summer months and at the turn of the year. This nature of the behavior is the same for other years, with obvious exceptions, which will be described in more detail in the next part of the text.
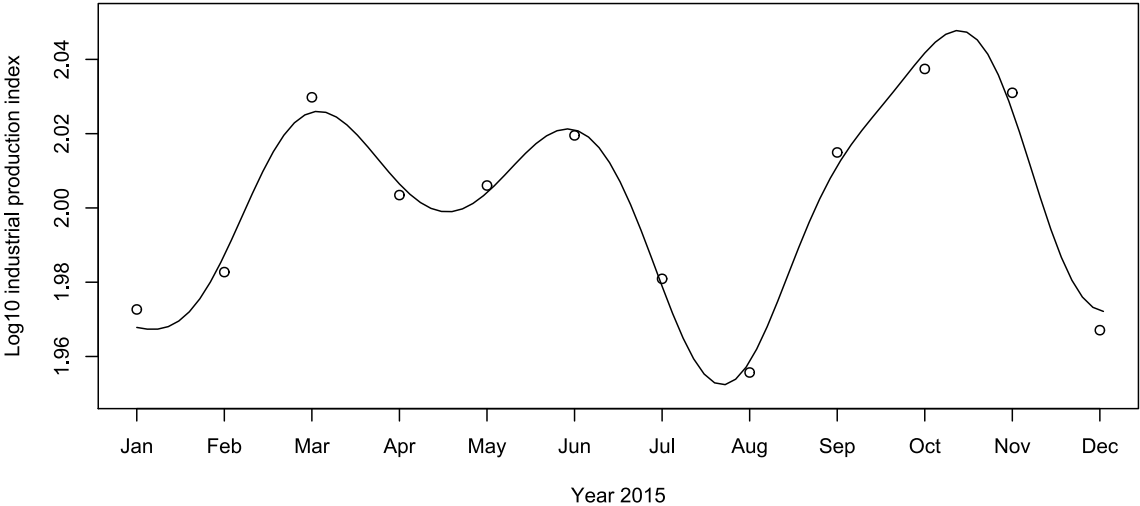
(a)
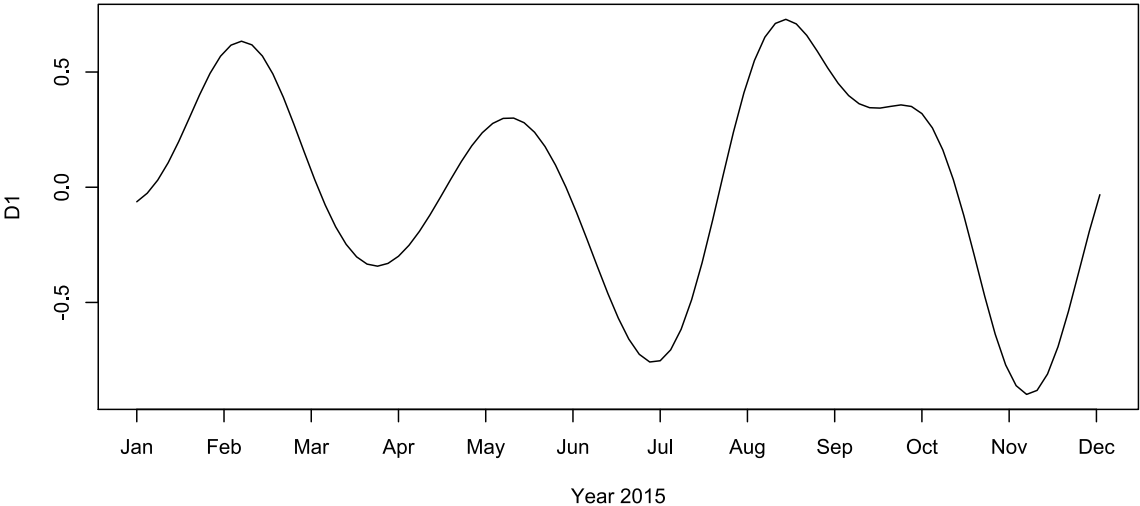
(b)
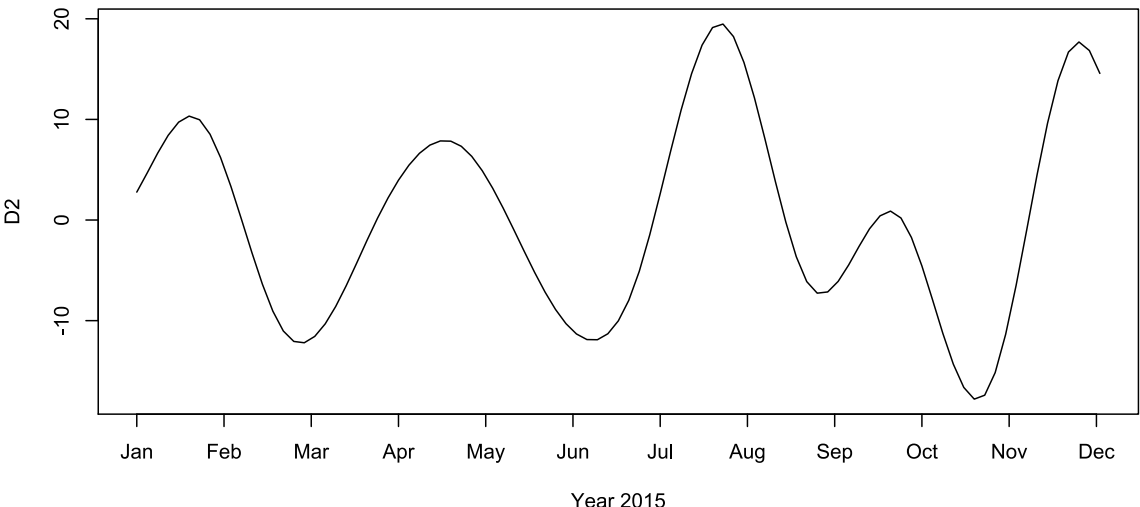
(c)

Figure 2. Estimated function (a) and its first (b) and second (c) derivative.

4.3. *Phase Plane Plots*

As already mentioned, one of the advantages of the functional approach is the possibility to work also with derivatives of functions. Usually, the first or the second derivative is used. In the case of functions of time, the first derivative expresses velocity, and the second derivative expresses acceleration. The so-called phase plane plot, in which the first and second derivatives are plotted against each other, can then serve as a useful analytical tool.

Phase plane plots are typically used in physics to visualize the behavior of oscillating systems, such as a mathematical pendulum, whose motion can be considered harmonic and described by a sinusoid. The system has a certain energy, which is periodically converted from the form of potential energy to the form of kinetic energy and vice versa. When the pendulum is tilted to the extreme position, the first derivative of the function of its position in time i.e. velocity is zero, so the kinetic energy is zero and the potential energy reaches a maximum. Conversely, when the pendulum is in a vertical equilibrium position, the speed of motion is maximum, acceleration i.e. the second derivative is zero, and thus the kinetic energy is maximum and the potential energy is zero. This energy transfer can be represented by a phase plane plot, see Figure 3, the motion of the pendulum is described by the function $\sin(2\pi t)$, the first derivative has the form $2\pi \cos(2\pi t)$ and the second derivative $-(2\pi)^2 \sin(2\pi t)$.
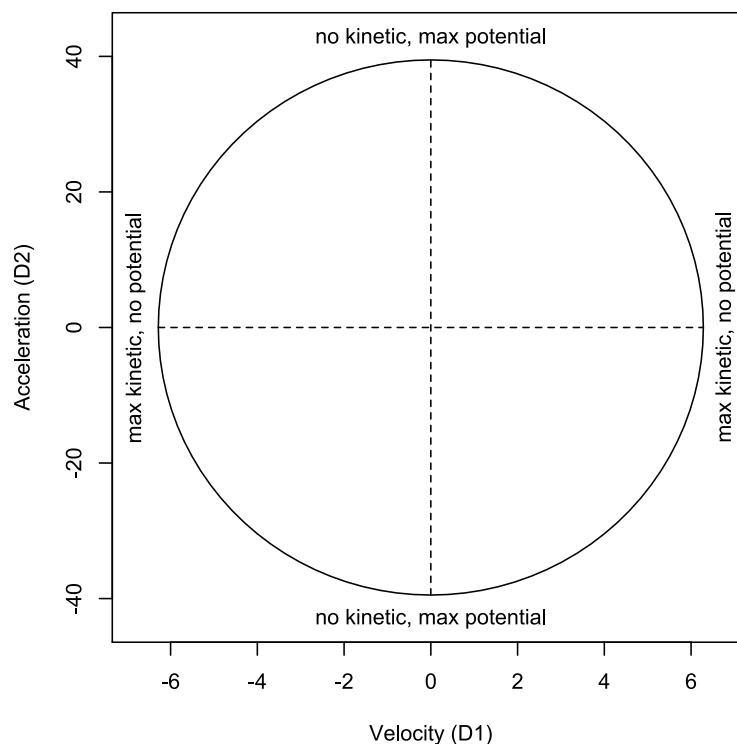

Figure 3. Phase plane plot for mathematical pendulum.

Energy transfer can be encountered also in other than just physical systems. From an economic point of view, available capital, human resources, and other factors which can be used, in the case of the index of industrial production, for industrial production, can be perceived as potential energy. The kinetic energy then corresponds to the ongoing production in which these resources are used. The production process undergoes periodic transitions, a period of high production (high kinetic energy), when resources are depleted (low potential

energy), is followed by a reduction in production (kinetic energy decreases) and replenishment of resources (potential energy increases), etc. In some periods, such as crises, however, there may also be situations where both the kinetic and potential energy are low, the phase plane curve is then close to zero.

4.4. Analysis of the Industrial Production Index

Phase plane plots were used for the analysis of industrial production in the Czech Republic in the period from 2000 to 2020. A phase plane plot was created for each year and the behavior of production in that year and the differences between years were examined.

Figure 4 shows a phase plane plot for the year 2015, which can be considered a typical year in a period of economic growth. The plot consists of a cyclic curve that begins near the middle with the designation of the month of January and continues clockwise throughout the year, with the labels indicating the months of the year. At the beginning of the year, the kinetic and the potential energy is low, during January the kinetic and the potential energy begins to grow, production increases, the rate of production growth first increases, then begins to decline, but production increases until March, when the rate drops to zero and then production will reduce. During April, production begins to grow again and continues to grow until June, when it begins to decline, but the decline is much more significant, which corresponds to the summer months and holidays, kinetic energy is low at this time and potential energy is high. In late summer, production increases rapidly, kinetic energy reaches a maximum and potential energy falls to zero, growth gradually slows down, production reaches its maximum in October and then begins to decline towards the end of the year.
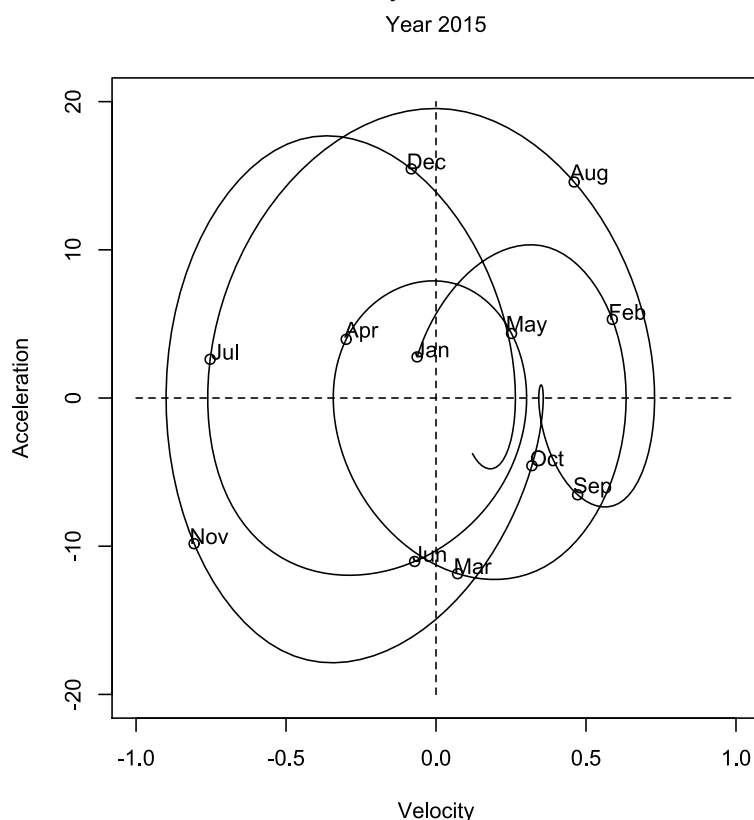

Figure 4. Phase plane plot for industrial production in the year 2015.

The characteristics of the development of industrial production are similar in other years, with the exception of the years when some significant events affecting the economy occurred. For comparison, the years 2008 and 2009, which were a period of economic crisis, as well as the year 2020 affected by the coronavirus pandemic will be shown.

Figures 5 and 6 show the phase plane plots for the years 2008 and 2009. In 2008, the beginning of the economic crisis is evident, in the second half of the year there is much less energy in the system than in other years and the usual significant autumn increase in production is lower and the decline continues until the following year. In the first half of 2009 there are more significant fluctuations in the production, the spring decline in production is more significant than in other years and the middle of the spring cycle is not shifted to the right to positive values of velocity corresponding to prevailing growth.

Figure 7 shows the phase plane plot for the year 2020. In March and April of this year, due to the onset of the coronavirus pandemic and the associated restrictions, a very significant decrease in production is evident (the plot is half the scale of previous plots). Subsequently, there is an increase in production, which peaks in June. The holiday decline in production is smaller than is typical in other years. The next second half of the year is already similar to the usual nature of production.
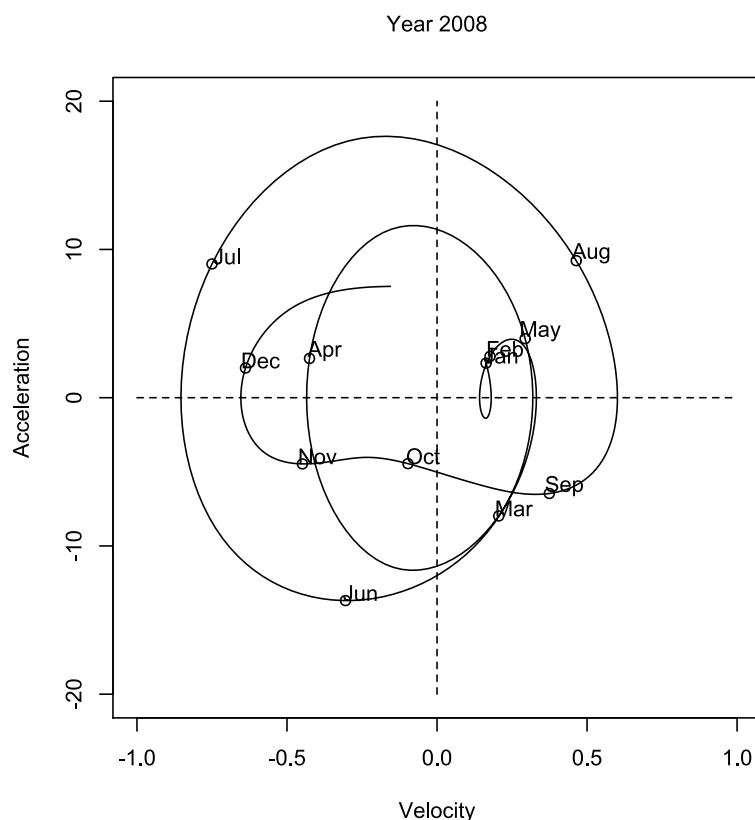

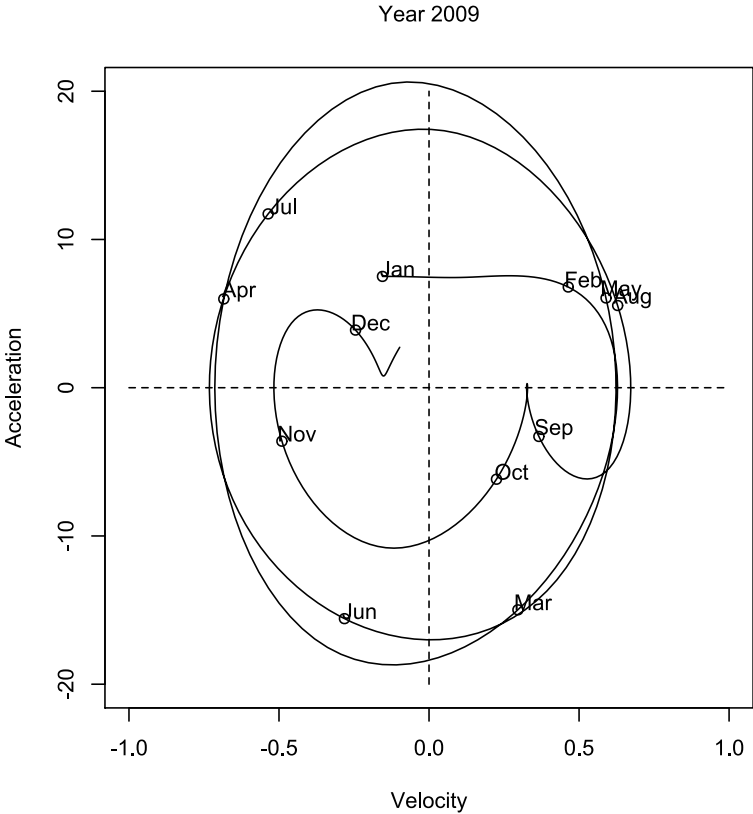**Figure 5.** Phase plane plot for industrial production in the year 2008.

Year 2009

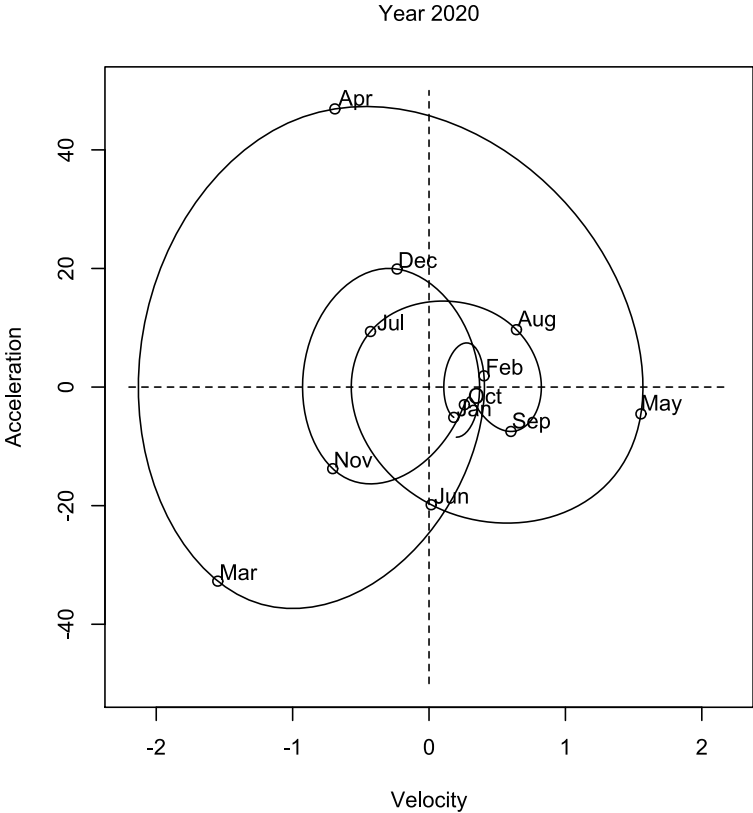**Figure 6.** Phase plane plot for industrial production in the year 2009.

Year 2020

Figure 7. Phase plane plot for industrial production in the year 2020.

5. Conclusions

The paper dealt with the possibilities of using functional data analysis in econometrics. A lot of economic data takes the form of observations of time-continuously evolving processes and can thus naturally be perceived as functional data. The search and review of articles showed that functional data analysis can find various applications in describing and modeling the behavior of various economic variables. Furthermore, the paper presented the main theoretical fundamentals of the functional approach and then showed its possible use in the exploratory analysis of the index of industrial production in the Czech Republic in the years 2000–2020. A functional representation of the data was created, and the derivatives were used in the analysis, which were plotted in phase plane plots. The plots were created for each year, which made it possible to study the seasonal character of production and compare differences between years, specifically focusing on the economic crisis in 2008 and 2009 and the coronavirus crisis in 2020. Phase plane plots allow better examination of the dynamics of industrial production, not only in terms of its growth or decline, but because of the representation of derivatives, it is possible to focus more on examining the rate of growth or decline and its acceleration or deceleration also with the possibility to observe the impact of various economic shocks. Interpretation with respect to energy transfer in the economic system is also possible. The exploratory analysis could be further followed, for example, by creating a functional model for the development of the industrial production index. The functional approach has other possibilities of use, working with derivatives of functions is just one of the benefits it brings. Functional data analysis can often help answer questions that are difficult for classical statistical methods.

References

Bapna, R., Jank, W., & Shmueli, G. (2008). Price formation and its dynamics in online auctions. *Decision Support Systems*, *44*(3), 641–656. https://doi.org/10.1016/j.dss.2007.09.004

Brockhaus, S., Fuest, A., Mayr, A., & Greven, S. (2018). Signal regression models for location, scale and shape with an application to stock returns. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *67*(3), 665–686. https://doi.org/10.1111/rssc.12252

Cai, T. (2018). Financial risk management based on functional data analysis. *Journal of Discrete Mathematical Sciences and Cryptography*, *21*(6), 1397–1400. https://doi.org/10.1080/09720529.2018.1527484

Caldeira, J., & Torrent, H. (2017). Forecasting the US Term Structure of Interest Rates Using Nonparametric Functional Data Analysis: Forecasting the US Term Structure of Interest Rates Using Nonparametric Functional Data Analysis. *Journal of Forecasting*, *36*(1), 56–73. https://doi.org/10.1002/for.2414

Caldeira, J., Gupta, R., Suleman, M. T., & Torrent, H. S. (2020). Forecasting the Term Structure of Interest Rates of the BRICS: Evidence from a Nonparametric Functional Data Analysis. *Emerging Markets Finance and Trade*, 1–18. https://doi.org/10.1080/1540496X.2020.1808458

Canale, A., & Vantini, S. (2016). Constrained functional time series: Applications to the Italian gas market. *International Journal of Forecasting*, *32*(4), 1340–1351. https://doi.org/10.1016/j.ijforecast.2016.05.002

Cao, R., Horváth, L., Liu, Z., & Zhao, Y. (2020). A study of data-driven momentum and disposition effects in the Chinese stock market by functional data analysis. *Review of Quantitative Finance and Accounting*, *54*(1), 335–358. https://doi.org/10.1007/s11156-019-00791-x

Czech Statistical Office. (2021). *Industry - Time series* [Data set]. Retrieved January 8, 2021 from https://www.czso.cz/csu/czso/pru_ts

Das, S., Demirer, R., Gupta, R., & Mangisa, S. (2019). The effect of global crises on stock market correlations: Evidence from scalar regressions via functional data analysis. *Structural Change and Economic Dynamics*, *50*, 132–147. https://doi.org/10.1016/j.strueco.2019.05.007

Hays, S., Shen, H., & Huang, J. Z. (2012). Functional dynamic factor models with application to yield curve forecasting. *The Annals of Applied Statistics*, *6*(3), 870–894. https://doi.org/10.1214/12-AOAS551

Hsing, T., & Eubank, R. L. (2015). *Theoretical foundations of functional data analysis, with an introduction to linear operators*. John Wiley and Sons, Inc.

Kearney, F., Cummins, M., & Murphy, F. (2018). Forecasting implied volatility in foreign exchange markets: A functional time series approach. *The European Journal of Finance*, *24*(1), 1–18. https://doi.org/10.1080/1351847X.2016.1271441

Kearney, F., Cummins, M., & Murphy, F. (2019). Using extracted forward rate term structure information to forecast foreign exchange rates. *Journal of Empirical Finance*, *53*, 1–14. https://doi.org/10.1016/j.jempfin.2019.05.002

Kearney, F., Murphy, F., & Cummins, M. (2015). An analysis of implied volatility jump dynamics: Novel functional data representation in crude oil markets. *The North American Journal of Economics and Finance*, *33*, 199–216. https://doi.org/10.1016/j.najef.2015.04.006

Kearney, F., & Shang, H. L. (2020). Uncovering predictability in the evolution of the WTI oil futures curve. *European Financial Management*, *26*(1), 238–257. https://doi.org/10.1111/eufm.12212

Kim, J.-M., & Jung, H. (2018a). Time series forecasting using functional partial least square regression with stochastic volatility, GARCH, and exponential smoothing: Functional partial least square regression. *Journal of Forecasting*, *37*(3), 269–280. https://doi.org/10.1002/for.2498

Kim, J.-M., & Jung, H. (2018b). Relationship between oil price and exchange rate by FDA and copula. *Applied Economics*, *50*(22), 2486–2499. https://doi.org/10.1080/00036846.2017.1400652

Kokoszka, P., & Reimherr, M. (2017). *Introduction to functional data analysis*. CRC Press, Taylor & Francis Group.

Kokoszka, P., Miao, H., & Zheng, B. (2017). Testing for asymmetry in betas of cumulative returns: Impact of the financial crisis and crude oil price. *Statistics & Risk Modeling*, *34*(1–2). https://doi.org/10.1515/strm-2016-0010

Kokoszka, P., Miao, H., Stoev, S., & Zheng, B. (2019). Risk Analysis of Cumulative Intraday Return Curves. *Journal of Time Series Econometrics*, *11*(2). https://doi.org/10.1515/jtse-2018-0011

Müller, H.-G., Sen, R., & Stadtmüller, U. (2011). Functional data analysis for volatility. *Journal of Econometrics*, *165*(2), 233–245. https://doi.org/10.1016/j.jeconom.2011.08.002

Oomen, R. (2019). Price signatures. *Quantitative Finance*, *19*(5), 733–761. https://doi.org/10.1080/14697688.2018.1532102

Peng, J., Paul, D., & Müller, H.-G. (2014). Time-warped growth processes, with applications to the modeling of boom–bust cycles in house prices. *The Annals of Applied Statistics*, *8*(3), 1561–1582. https://doi.org/10.1214/14-AOAS740

Ramsay, J. O., & Silverman, B. W. (2005). *Functional data analysis* (2nd ed). Springer.

Reddy, S. K., & Dass, M. (2006). Modeling On-Line Art Auction Dynamics Using Functional Data Analysis. *Statistical Science*, *21*(2), 179–193. https://doi.org/10.1214/088342306000000196

Reithinger, F., Jank, W., Tutz, G., & Shmueli, G. (2008). Modelling price paths in on-line auctions: Smoothing sparse and unevenly sampled curves by using semiparametric mixed models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *57*(2), 127–148. https://doi.org/10.1111/j.1467-9876.2007.00605.x

Scott, M. A., & Handcock, M. S. (2005). Persistent Inequality? Answers From Hybrid Models for Longitudinal Data. *Sociological Methods & Research*, *34*(1), 3–30. https://doi.org/10.1177/0049124105277194

Seya, H., Yoshida, T., & Tsutsumi, M. (2016). Ex-post identification of geographical extent of benefited area by a transportation project: Functional data analysis method. *Journal of Transport Geography*, *55*, 1–10. https://doi.org/10.1016/j.jtrangeo.2016.07.004

Wang, D., Xu, Y., He, L., & Liu, R. (2018). Spatial and Temporal Differences in the Relationships between Residents' Income and Consumption in China: A Dynamic Analysis Using Functional Data Analysis. *Sustainability*, *10*(12), 4703. https://doi.org/10.3390/su10124703

Wang, S., Jank, W., & Shmueli, G. (2008). Explaining and Forecasting Online Auction Prices and Their Dynamics Using Functional Data Analysis. *Journal of Business & Economic Statistics*, *26*(2), 144–160. https://doi.org/10.1198/073500106000000477

Wang, Z., Sun, Y., & Li, P. (2014). Functional Principal Components Analysis of Shanghai Stock Exchange 50 Index. *Discrete Dynamics in Nature and Society*, *2014*, 1–7. https://doi.org/10.1155/2014/365204