Air Quality, Atmosphere & Health (2023) 16:25–36
https://doi.org/10.1007/s11869-022-01261-5

Links between the concentrations of gaseous pollutants measured in different regions of Estonia

Aare Luts[1] · **Marko Kaasik**[1,2] · **Urmas Hõrrak**[1] · **Marek Maasikmets**[2] · **Heikki Junninen**[1]

Received: 23 February 2022 / Accepted: 21 September 2022 / Published online: 14 October 2022
© The Author(s), under exclusive licence to Springer Nature B.V. 2022

Abstract
The factors that determine the concentrations of air pollutants (NO, NO_2, SO_2, O_3), measured in 8 monitoring stations (4 rural background, 3 urban, and 1 industrial) in Estonia, are studied applying the factor analysis. The factor analysis reveals remarkable impact of COVID-19 lockdown, effects caused by dramatic decrease in oil-shale based energy production in Estonia provoked by new socio-economic conditions such as elevated price for CO_2 emission quota, differences between rural and urban stations, maritime-continental difference for NO_2 and ozone, and specific industrial impact in case of SO_2. The multiple regression analysis to predict the ozone concentration in one rural background station at Tahkuse was performed, based on the ozone concentrations measured in other stations and the concentrations of NO, NO_2, and CO_2, recorded in the same station. It was found that the ozone concentration at Tahkuse is rather well predictable (determination coefficient, i.e., correlation coefficient squared, $R^2 = 0.714$), using only the concentrations from another rural station at Saarejärve that is about 110 km away from Tahkuse. Adding all the available data into the list of regression analysis arguments, the model predictability is improved moderately (determination coefficient $R^2 = 0.795$). Large model residuals above all tend to occur with the values measured and predicted at summer nights. Surprisingly, neither NO nor NO_2 concentration measured in the Tahkuse station did appear a good predictor for ozone ($R^2 = 0.02$ and 0.05, respectively), possibly long-range transport of ozone (that has also experienced NO and/or NO_2 influence during transport) overrides the local effects of NO and/or NO_2.

Keywords Nitrogen oxides · Sulphur dioxide · Ozone · Factor analysis · Multiple regression

Introduction

The composition and concentrations of trace (pollutant) gases and aerosol particles in the ambient air depend on many factors, e.g., on natural and anthropogenic (pollutant) emissions and transport, air chemistry, and planetary boundary layer mixing state. The links between these air pollutant concentrations and the links with environmental parameters are topics of extensive studies, e.g., Nogarotto and Pozza (2020). The review paper by Nogarotto and Pozza also evaluates the mathematical methods employed for such tasks. Regression models combining air pollution concentration with air mass trajectories, principal component analysis (PCA), and factor analysis are examples of the common methods.

Factor analysis enables to reveal the latent factors that determine the variations in the observed concentrations of a set of investigated variables. The latent factors may include meteorological parameters (humidity, temperature, wind, etc.) but also some other environmental parameters, like the concentrations of specific chemical compounds that may better help to identify the sources of the pollutants under interest. Shubhankar and Ambade (2016) studied spatial and temporal variation of patterns of ambient air pollutants. They concluded that three factors were responsible for the variability of most of the observed variables, whereas emissions from the transport and industrial output were among these factors. Chan and Mozurkewich (2007) determined the origins of the measured particulate matter using absolute principal component analysis. They revealed three to four common factors, e.g., long-range transport, and few additional factors that mostly characterize specific measurement site(s), e.g., local industry. The emphasis of the paper Chan

✉ Aare Luts
aare.luts@ut.ee

[1] Institute of Physics, Tartu University, Ostwaldi str. 1, Tartu, Estonia

[2] Estonian Environmental Research Centre, Tallinn, Estonia

and Mozurkewich (2007) is on the procedure development, and all the factors are not clearly identified. Tai-Yi Yu (2010) demonstrated the use of emission inventory, concentrations of ambient air pollutants, and PCA approach to provide new information about particulate matter PM2.5 concentrations. Four components out of 76 rotational components were cited as major factors. Liu et al. (2020) studied PM origin using the links with certain chemical compounds. The main five factors found out were secondary inorganic aerosol, coal combustion, crustal dust, vehicle exhaust, and biomass burning.

Several factors can be specific to certain particular measurement site. Studies of such specificity can be related to the problem of proper allocation of measurement sites (e.g., establishment of new measurement stations). In case the variations of investigated variables observed at some particular site cannot be described (simulated) by common factors, the measurement results obtained at this site contain a unique information. He and Lu (2012) employed multiple regression and principal component analysis to predict ozone levels at two sites from the data set on air pollutants and meteorological variables. They concluded that the ozone level depends on many parameters, but the ratio NO_2/NO_x explained 75% of variation of the ozone level at Sha Tin and 83% at Kwun Tong, Hong Kong, China. When they looked for the predicted ozone concentration at a certain site, the ozone concentrations measured at another site were not considered. Lu et al. (2011) and Araki et al. (2015) searched for an optimal location of the environmental monitoring stations, applying the criterion that they adequately represented the air pollutant concentrations in the domain of interest. They demonstrated that for each air pollutant, the monitoring stations could be grouped into different classes based on their air pollution patterns. Li et al. (2022) implemented a multifactor linear model to predict PM concentrations in Beijing region, including socioeconomic factors. Recently, several machine learning algorithms like online sequential extreme learning machine (OS-ELM that is in essence a modification of neural network) are tested to predict the distribution of polluting gases and particles, e.g., by Sharma et al. (2020). As an outcome, the results by OS-ELM for several datasets can yield better results that the ones obtained by traditional methods like multiple linear regression, but in other cases, the traditional methods are quite on a par with the new algorithms. Xu et al. (2022) evaluated several methods including PCA for reducing the dimensionality of the data to model spatial variation of gaseous air pollutants. They discovered that in several cases, the random forest method yielded the best results, in certain other cases different methods, including linear regression model performed better. Wang et al. (2020) employed a complex method for air quality index forecasting and discovered that a complex method that contains various signal processing and optimization methods, e.g., the Hampel identifier, variational mode decomposition (VMD), sine cosine algorithm (SCA), and extreme learning machine, yields better results than the simpler methods. Unfortunately, this method resembles a mathematical "black box" that is not easily interpretive and, in addition, this method is not openly available.

In this study, we aim to find out the links between the patterns of the atmospheric gas composition changes measured in various parts of Estonia, in background air monitoring stations and urban stations. We also search for a model that can predict the ozone concentration at one station. Besides new knowledge, such model makes it possible to fill the occasional gaps in the data with approximated values better than any interpolations can do. Filling in the data gaps, in turn, facilitates to calculate certain proxies that depend on the ozone concentrations.

The results of this study add new knowledge about the (latent) factors that determine selected trace gas concentration variations in Estonia and the ability of factor analysis to reveal such relationships in the cases that study the spatial–temporal distribution of trace gas concentration variations. Additionally, the results clarify the influence of the restrictions enforced during COVID pandemic. The outcomes of the comparison of several regression methods demonstrate that in certain cases, the linear methods can be considered optimal. The elaborated model that can predict the ozone concentration once more demonstrates the typical parts of common and local factors, enables to fill the occasional gaps in the data with approximated values, and also points out certain guidelines about directions of future studies.

The paper is organized as follows. First, we introduce the used data sources likewise the used methods. The latter part includes the results of comparison of certain regression methods. Then, we discuss the concentration variation patterns of atmospheric gases and the discovered latent factors that determine the patterns. Next, we introduce and discuss the elaborated regression model that can predict the ozone concentration variations in the one measurement station (Tahkuse). Finally, we integrate the outcomes by discussion and summary.

Methods

We have used the measured about 6-year (from 01.01.2016 to 30.09.2021) data on hourly averaged concentrations of atmospheric trace and polluting gases O_3, NO, NO_2, and SO_2 from the Estonian national air quality monitoring stations: background stations Lahemaa, Vilsandi (maritime, on a small island), and Saarejärve; urban stations Tallinn-Liivalaia, Tallinn-Õismäe, and Tartu; and industrial station Kohtla-Järve (all these stations are

operated by Estonian Environmental Research Centre), and also the data from the Tahkuse rural air monitoring station operated by the University of Tartu. All above mentioned monitoring stations are contributing to the national air quality monitoring system (http://õhuseire.ee/en). The data about concentrations have been measured using the gas analysers and the procedures commonly approved in the field of routine environmental gas composition monitoring. The procedures include regular calibration and maintenance of the devices. Gas analysers used in monitoring stations are listed in Tables 1 and 2. Available time resolution refers to data collected by national air quality monitoring system.

In the time series for factor analysis and regression analysis, at each particular hour, the data from all the stations must be present; otherwise, this particular hour remains out of analysis. Within the whole nearly 6-year time interval, about 85% from all hours met this data integrity criteria.

Entire time series of measurements 2016–2021 has been divided into two sub-periods of 2016–2019 and 2020–2021 to find out the possible effects of COVID-19 pandemic–related restriction measures to atmospheric air quality. Also, since late 2019, operation of oil-shale-fired thermal power plants was almost suspended due to high price of CO_2 quota, thus making the secondary difference in emissions for 2020–2021.

We performed factor analysis (Varimax rotated) to search for the latent factors that determine the gas concentrations in those stations. Factor analysis is a statistical data reduction technique used to explain observed variations among observed variables (e.g., the concentrations measured at specific stations) in terms of fewer new variables named factors (e.g., long-range transport factor that can explain certain part of observed variations). Varimax rotation is a statistical technique used to clarify the relationship among factors. It is intended to maximize the variance shared among items. By maximizing the shared variance, results more discretely represent how data correlate with each factor component. The measurement sites of monitoring network are considered representative for certain pollution types: rural background in various parts of the country at different distances from the sea (Tahkuse, Saarejärve, Lahemaa, Vilsandi), urban-industrial (Kohtla-Järve), urban background (Tallinn-Õismäe, Tartu), and urban street (Tallinn-Liivalaia). Therefore, the combination of the factors should reveal the processes that determine the air pollution pattern in Estonia.

We also apply the regression analysis techniques to model and predict the ozone concentration at Tahkuse station, using the measured data from the other monitoring stations surrounding Tahkuse, as independent variables. One application of such a model is to fill gaps within the time series of the measurement results. There are several known techniques for that caps-filling task, discussed e.g. by Junninen et al. (2004), but none of these methods performs best in all the cases. Now, we can use the data from other (nearby) stations; therefore, the first choice was to use these data to build a model for predicting the values at Tahkuse, not just to try fill the gaps in the data according the methods discussed by Junninen et al. (2004). The predicted results are compared to original one recorded at Tahkuse to estimate the quality of used model. This way, we can assess the common and the specific components within the ozone concentration variations.

Table 1 Gas analyzers used in Lahemaa, Vilsandi, Saarejärve, Liivalaia (Tallinn), Õismäe (Tallinn), Tartu, and Kohtla-Järve monitoring stations

Measured quantity	Instrument	Company	Available time resolution
NO_X, NO, and NO_2	APNA-360 ambient NOx monitor	Horiba Ltd	30 min
SO_2	APSA-360 ambient sulfur dioxide monitor	Horiba Ltd	30 min
O_3	APOA-360 ambient ozone monitor	Horiba Ltd	30 min

Table 2 Gas analyzers of Tahkuse air monitoring station

Measured quantity	Instrument	Company	Available time resolution
NO_X, NO, and NO_2	Model 42i-TL TRACE Level NOx Analyzer	Thermo Scientific™	30 min
SO_2	43i-TLE enhanced trace level SO_2 analyzer	Thermo Scientific™	30 min
O_3	Model 49i ozone analyzer	Thermo Scientific™	30 min
CO_2, CH_4, H_2O	Model 911–0010 Greenhouse Gas Analyzer FGGA-24EP	Los Gatos Research	30 min

Table 3 Comparison of the RMSE (root-mean-square error) values obtained by several methods

Combination	MLR	Neural network	Decision tree	SVM
Model a training	11.84	11.85	12.27	11.67
Model a test with a	11.84	11.71	12.72	12.10
Model a test with b	10.96	10.82	11.79	11.03
Model b training	10.63	10.69	11.22	10.61
Model b test with b	10.69	10.54	11.03	10.59
Model b test with a	12.2	12.24	12.48	11.99

Regression model can be built by several methods: multiple linear regression abbreviated as MLR, neural networks, decision trees, support vector machines abbreviated as SVM, etc. We have tested these methods, and a summary of the comparison results is presented in Table 3. The methods are trained and implemented separately for subperiods of 2016–2019 (period a) and 2020–2021 (period b) and evaluated both by the test data sets taken randomly from the "own" subperiod (e.g., "Model a test with a") and by the data from the other subperiod (e.g., "Model a test with b").

The model trained by the data taken from "period a" is marked by "model a"; the one trained by the data from "period b" is marked by "model b." The names of the combinations in Table 3 include both the model name and the name of the period where the test data are taken from, e.g., "model a test with b" means that the model a is tested with the data taken from the period b.

According to the RMSE (root-mean-square error) values, two better methods are multiple linear regression and neural network. We tested several versions of neural network, and the best one among these included three layers with ten nodes in every layer and node activation by RELU (rectified linear unit) function. Decision tree models demonstrated the worst RMSE values. The results achievable by SVM models depended on the dataset. These models are also rather hard to interpret; therefore, we omitted the SVM models from further study. Then, we carried out a bootstrapping analysis using random datasets from the whole time series to train and test the linear and neural network models. According to this analysis, the differences in the RMSE values obtained by linear and neural network models are statistically relevant at all test cases, whereas sometimes neural network performs better. Nevertheless, the differences in the RMSE values (although statistically relevant) are not large (within ca one per cent from the corresponding RMSE values), and the predicted *vs* measured scatterplots look rather similar for both methods. Whereas the training of a neural network is up to several tens of times slower and a linear model is easier to interpret, we selected multiple linear regression model for subsequent analysis.

Table 4 The scores of the first five factors that determine NO concentration variations at specific stations (2016–2019) and the determination powers of the factors

Station	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Kohtla-Järve NO	*−0.51*	0.10	0.06	−0.26	−0.11
Õismäe NO	*−0.46*	−0.11	0.00	0.03	−0.03
Liivalaia NO	*−0.40*	−0.08	−0.02	0.00	0.01
Tartu NO	0.13	−0.03	−0.01	*1.02*	−0.01
Saarejärve NO	−0.01	*0.38*	0.05	−0.09	0.11
Tahkuse NO	0.10	*0.85*	0.03	0.00	−0.19
Lahemaa NO	0.09	−0.10	0.06	0.01	*1.01*
Vilsandi NO	0.01	−0.04	*−1.01*	0.00	−0.07
Factor power, %	29.2	16.0	11.9	10.7	9.6
Cumulative power, %	29.2	45.2	57.0	67.7	77.2

Results

The NO concentration variation patterns

The results of factor analysis are presented in Tables 4 and 5.

All the factor scores are calculated by using the Real Statistics package (2021). When compared with the other studied trace gases (NO_2, SO_2, and ozone), the NO concentrations are most location-specific. This feature can be understood as a result of rapid chemical conversion of NO, before advected at long distances. In the period 2016–2019, the concentrations at Kohtla-Järve, Õismäe, and Liivalaia stations behave remarkably similarly, but in the period 2020–2021, the same is valid for the concentrations at Õismäe, Liivalaia, and Tartu. In the earliest period, the concentrations at Tartu vary in different way; in the later period, the variations at Kohtla-Järve are different. Saarejärve and Tahkuse sites

Table 5 The scores of the first five factors that determine NO concentration variations at specific stations (2020–2021) and the determination powers of the factors

Station	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Kohtla-Järve NO	0.13	0.00	0.03	−0.01	*−1.03*
Õismäe NO	*−0.50*	−0.06	0.01	0.00	0.05
Liivalaia NO	*−0.40*	−0.06	−0.04	0.00	0.03
Tartu NO	*−0.39*	0.06	0.05	−0.04	0.14
Saarejärve NO	−0.05	*0.44*	−0.03	−0.03	0.03
Tahkuse NO	0.11	*0.82*	0.10	−0.06	0.01
Lahemaa NO	0.03	−0.04	0.05	*1.01*	0.01
Vilsandi NO	0.03	−0.01	*−1.00*	−0.04	0.03
Factor power, %	26.8	15.6	12.1	11.4	10.1
Cumulative power, %	26.8	42.4	54.5	65.9	76.5

Table 6 The scores of the first five factors that determine NO_2 concentration variations at specific stations (2016–2019) and the determination powers of the factors

Station	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Kohtla-Järve NO_2	0.19	−0.25	0.00	0.04	*−1.10*
Õismäe NO_2	*−0.45*	−0.06	0.02	−0.16	−0.05
Liivalaia NO_2	*−0.80*	−0.05	0.01	−0.16	0.33
Tartu NO_2	0.15	−0.02	−0.01	*1.05*	0.11
Saarejärve NO_2	0.12	*0.49*	−0.11	0.08	0.13
Tahkuse NO_2	0.00	*0.47*	−0.05	0.11	0.29
Lahemaa NO_2	−0.02	*0.43*	−0.10	*−0.50*	−0.26
Vilsandi NO_2	−0.01	−0.15	*1.05*	0.00	0.01
Factor power, %	42.2	17.8	9.9	7.8	7.0
Cumulative power, %	42.2	60.0	69.9	77.7	84.7

Table 7 The scores of the first five factors that determine NO_2 concentration variations at specific stations (2020–2021) and the determination powers of the factors

Station	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Kohtla-Järve NO_2	0.12	0.15	−0.01	−0.11	*−1.11*
Õismäe NO_2	0.08	*−0.42*	0.06	−0.15	−0.11
Liivalaia NO_2	0.03	*−0.81*	0.01	−0.12	0.32
Tartu NO_2	0.16	0.16	0.00	*1.14*	0.10
Saarejärve NO_2	*−0.49*	0.14	−0.18	−0.01	0.01
Tahkuse NO_2	*−0.55*	−0.02	−0.12	0.03	0.31
Lahemaa NO_2	*−0.36*	−0.01	−0.08	−0.18	0.00
Vilsandi NO_2	0.21	−0.03	*1.09*	−0.02	−0.01
Factor power, %	42.8	18.8	8.6	8.2	7.3
Cumulative power, %	42.8	61.6	70.2	78.4	85.7

Table 8 The scores of the first five factors that determine SO_2 concentration variations at specific stations (2016–2019) and the determination powers of the factors

Station	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Kohtla-Järve SO_2	−0.01	−0.04	*−1.00*	−0.02	0.00
Õismäe SO_2	*0.56*	0.07	−0.02	−0.05	−0.11
Liivalaia SO_2	*0.59*	0.10	0.00	−0.01	−0.15
Tartu SO_2	0.03	0.17	−0.01	*0.91*	−0.07
Saarejärve SO_2	−0.19	−0.23	0.08	0.28	0.21
Tahkuse SO_2	−0.10	*−0.51*	−0.01	0.03	−0.01
Lahemaa SO_2	-0.10	0.11	0.01	−0.02	*1.04*
Vilsandi SO_2	−0.03	*−0.62*	−0.03	−0.21	−0.13
Factor power, %	33.8	15.1	12.9	11.3	9.4
Cumulative power, %	33.8	49.0	61.9	73.2	82.6

Table 9 The scores of the first five factors that determine SO_2 concentration variations at specific stations (2020–2021) and the determination powers of the factors

Station	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Kohtla-Järve SO_2	0.05	0.04	0.03	*−1.00*	0.00
Õismäe SO_2	*0.49*	0.04	−0.03	−0.03	−0.05
Liivalaia SO_2	*0.56*	0.05	0.00	0.00	−0.21
Tartu SO_2	−0.15	*−0.79*	−0.13	0.00	0.05
Saarejärve SO_2	0.03	*−0.48*	0.09	0.05	−0.17
Tahkuse SO_2	−0.07	0.03	*1.01*	−0.02	−0.06
Lahemaa SO_2	0.16	0.05	−0.05	−0.03	0.16
Vilsandi SO_2	−0.12	0.09	−0.07	−0.01	*1.02*
Factor power, %	31.6	14.2	12.7	11.4	9.6
Cumulative power, %	31.6	45.8	58.5	69.9	79.5

constitute another group that demonstrates somewhat analogous features. The concentrations at both Vilsandi and Lahemaa vary clearly in their own way.

The NO_2 concentration variation patterns

NO_2 concentrations are essentially determined by three factors (Tables 6 and 7). The NO_2 concentrations at rural stations Saarejärve, Tahkuse, and Lahemaa demonstrate similar variations both in the period 2016 – 2019 (Table 6, factor 2) and also in the period 2020–2021 (Table 7, factor 1).

There are certain differences in the particular variations during both sub-periods, too, but within the less essential factors 4 and 5. The Tallinn Liivalaia station in the centre of the city and the urban background station Õismäe form another joint group according to the factor scores. Liivalaia station, a traffic-affected site, has a stronger impact onto the factors. According to the factors 1–3, the concentrations at Tartu and Kohtla-Järve show certain common features (rather small, but similar factor scores), but they differ by factors 4 and 5. The Vilsandi station demonstrates clearly distinctive variations in the concentrations. The factor 4 (weight about 8%) shows some anticorrelation considering Tartu city station and Lahemaa background air monitoring station data during the period 2016–2019. During the period 2020–2021 (Table 7), the contribution of Lahemaa NO_2 to factor scores is weaker in general, and beside the factor 1, it is showing very weak or insignificant effect. During the period 2020–2021 (Table 7), the factor related to NO_2 variations recorded in background air monitoring stations became to the first place, while it was in the second position in the period 2016–2019. Vilsandi background air monitoring station, located in the Vilsandi Island differs from other background stations, probably due to well-expressed marine conditions.

The SO_2 concentration variation patterns

The main results of factor analysis are presented in Tables 8 and 9. The SO_2 concentrations are remarkably determined by the first two factors, but the site-specific factors are important, too. However, factors 3–5 do not describe much more than the SO_2 measured in each of the stations highly contributing to those factors, because these factors contain only one large factor score.

Firstly, Tallinn urban stations Liivalaia and Õismäe can be grouped together according to the similar variations in SO_2 concentrations. The similarity of variations in other stations depends on the time period. During the period 2016–2019, the concentration variations at Tahkuse and Vilsandi background stations show common features (factor 2), but during the period 2020–2021, these variations are clearly distinct each other according the factors 3 and 5. The variations at Kohtla-Järve station have clearly individual character; mostly the same is valid also for the Lahemaa station, especially for the period 2016–2019.

The ozone concentration variation patterns

Ozone concentrations are essentially determined by the first factor. According to the factor analysis, the most important first factor accounts the nearly equal contribution from all stations; the local individualities of stations become evident within the next factors (Tables 10 and 11). The behaviour of ozone concentration at the monitoring stations can be allocated to a few more closely bounded groups, and also to certain groups where the links are weaker, the groups are mentioned below. The links between the stations also depend on the time period. During both periods, the variations at Liivalaia and Vilsandi stations are the most essential terms within factors 2 and 3, even though within the factor 3 the impact of Liivalaia station is opposite to the one of Vilsandi station. During 2020–2021, the variations in Tartu and Tahkuse stations are the most important terms of the factor 4 with effects that have the opposite signs. During 2016–2019, only the variations in Tartu dominate within the factor 4. Factor 5 is largely determined by the variations at Kohtla-Järve station. During the period 2016–2019, this factor also contains large part that is induced only by the variations at Tahkuse station, whereas during the period 2020–2021, the second and the third roles within this factor are induced by the variations at both Tartu and Tahkuse stations.

Table 10 The scores of most important factors that determine ozone concentration variations at specific stations (2016–2019) and the determination powers of the factors

Station	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Kohtla-Järve ozone	−0.19	−0,39	−0.19	*−0.48*	*−1.05*
Õismäe ozone	−0.12	0.34	−0.30	−0.19	−0.03
Liivalaia ozone	−0.15	*0.91*	*−0.64*	0.10	0.28
Tartu ozone	−0.15	−0.28	0.02	*1.63*	−0.03
Saarejärve ozone	−0.11	−0.36	0.01	−0.05	−0.18
Tahkuse ozone	−0.17	−0.40	0.17	−0.41	*1.63*
Lahemaa ozone	−0.14	−0.20	−0.16	−0.45	−0.30
Vilsandi ozone	−0.12	*0.52*	*1.25*	−0.09	−0.29
Factor power, %	72.9	7.6	6.2	5.0	3.8
Cumulative power, %	72.9	80.4	86.6	91.6	95.4

Table 11 The scores of most important factors that determine ozone concentration variations at specific stations (2020–2021) and the determination powers of the factors

Station	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Kohtla-Järve ozone	−0.17	−0.31	−0.14	−0.25	*−1.24*
Õismäe ozone	−0.13	0.29	−0.41	0.20	−0.07
Liivalaia ozone	−0.15	*0.76*	*−0.85*	0.22	0.39
Tartu ozone	−0.15	−0.28	−0.02	*−1.44*	*0.92*
Saarejärve ozone	−0.12	−0.41	0.13	0.16	−0.05
Tahkuse ozone	−0.17	−0.43	0.38	*1.05*	*0.97*
Lahemaa ozone	−0.14	−0.23	−0.13	0.20	*−0.63*
Vilsandi ozone	−0.12	*0.74*	*1.19*	−0.18	−0.24
Factor power, %	76.7	7.3	5.0	3.7	3.2
Cumulative power, %	76.7	83.9	88.9	92.7	95.9

The model to predict the ozone concentrations at Tahkuse

We made attempts to estimate the ozone concentrations at Tahkuse by multiple linear regression method using the data from Tahkuse as dependent variable and from other stations as independent variables. We also added the concentrations of NO, NO_2, and CO_2, measured in Tahkuse station, into the list of independent arguments and studied their impact onto the power of the model to predict the actual ozone concentrations observed in Tahkuse station. We performed the analysis separately for two periods: for years 2016–2019 and for 2020–2021. When to consider the simple regressions with only one argument, the highest coefficient of determination R^2 corresponds to the regression where the ozone data from Saarejärve are used as independent variable: $R^2 = 0.677$ for the period 2016–2019 and $R^2 = 0.714$ for the period 2020–2021. In the case of the multiple linear regression, where the data from Vilsandi and Lahemaa were added to Saarejärve, the values of coefficient R^2 become equal to 0.739 for the period 2016–2019 and 0.754 for the period 2020–2021. Tahkuse and Saarejärve are the two most

continental rural background stations in Estonia. Vilsandi represents relatively clean environment and marine conditions; the Lahemaa station (6 km from the coast of Gulf of Finland) is another site considerably affected by coastal mesoscale weather patterns.

In the case of the multiple regression, where the ozone data from all the stations (Saarejärve, Lahemaa, Vilsandi, Kohtla-Järve, Õismäe, Liivalaia, and Tartu) and also the NO, NO_2, and CO_2 concentrations (CO_2 in ppm, other gases in μg m^{-3}) measured at Tahkuse station were used as independent variables, the determination coefficients $R^2 = 0.767$ for the period 2016–2019 and 0.795 for the period 2020–2021, whereas without the CO_2 concentrations, the R^2 values are 0.745 and 0.762, respectively. When using the NO and/or NO_2 concentrations as sole independent arguments, the determination coefficients R^2 stay below 0.05; therefore, we can leave these concentrations out from the list of the parameters of the model. Using only CO_2 concentrations in the role of independent argument enables to yield the determination coefficient R^2 values up to 0.26. Inclusion of the CO_2 concentrations to the list of independent arguments in addition to ozone data from the other measurement stations certainly enhances the R^2 values. Unfortunately, in these cases, certain unpleasant side effects appear: such a model too often predicts negative ozone concentrations (usually in the case of high CO_2 concentrations recorded in summer nights with low winds), which is physically intolerable. For that reason, we omit the CO_2 concentrations from the list of independent arguments, despite the fact that CO_2 can numerically enhance the R^2 values by about 2–3%. As concentration of ozone has obvious diurnal pattern, we also tried to include time as an independent parameter, but without any remarkable success. Below we discuss the model with argument list that contains the concentrations of ozone measured in Saarejärve, Vilsandi, and Lahemaa stations. In this case, the determination power of the model is almost so good as in the case of longer lists of arguments. Only CO_2 can enhance the R^2 values by some extent, but it was excluded by the physical reasons as described above.

The multiple regression models are different for the periods 2016–2019 (Eq. 1) and 2020–2021 (Eq. 2). Below we also discuss the results obtained when Eq. 1 was applied to the period 2020–2021 and when Eq. 2 was applied to the period 2016–2019.

$$Y = 0.4635C_1 + 0.3455C_2 + 0.1797C_3 + 0.6139 \quad (1)$$

$$Y = 0.5806C_1 + 0.2071C_2 + 0.2034C_3 + 2.634 \quad (2)$$

where C_1, C_2, and C_3 are the concentrations measured in Saarejärve, Lahemaa, and Vilsandi, respectively, and Y is the predicted concentration for Tahkuse station (μg m^{-3}).

The ozone prediction model results for Tahkuse station are depicted in Fig. 1 (Eqs. 1 and 2 applied to the period 2016–2019) and Fig. 2 (Eqs. 1 and 2 applied to the period 2020–2021). Figures 1 and 2 contain the colour plots (predicted concentrations vs measured ones) and the trendlines with the trendline equations shown. The summary statistics of comparison of the models is presented in Table 12. The first column contains list of the parameters that can be used to evaluate the model (Willmott et al., 1985). RMSE is root-mean-square error, RMSE_s is systematic part of root

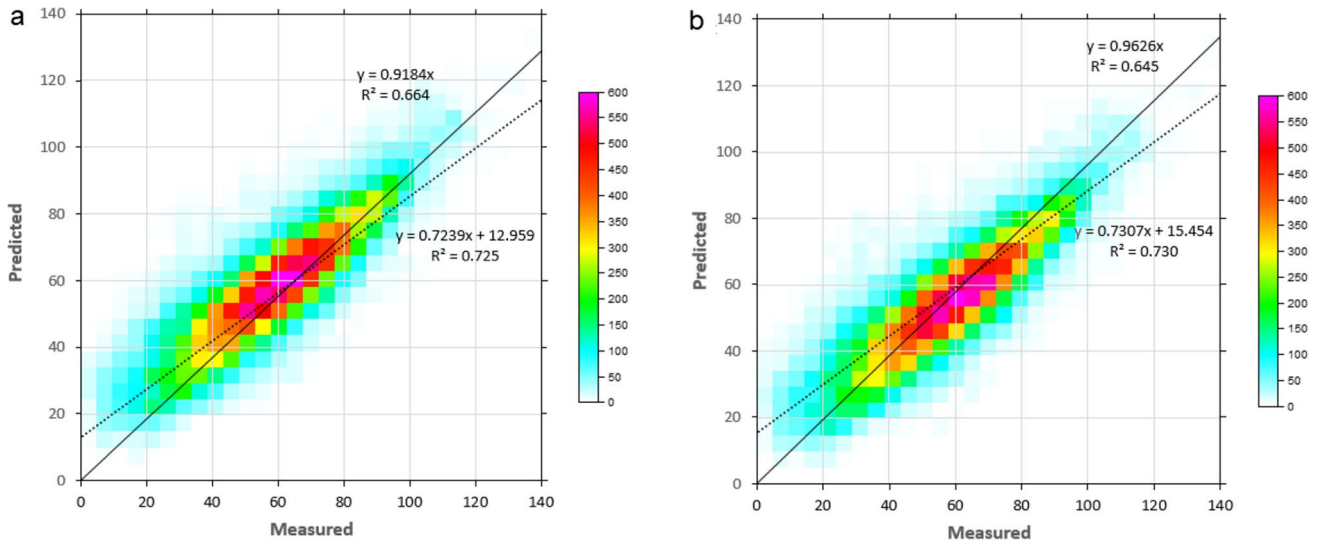

Fig. 1 Concentrations of ozone (μg m^{-3}) for years 2016–2019: **a** predicted from regression Eq. 1 and **b**) predicted from regression Eq. 2. Colour scale represents the number of cases in 5 by 5 μg m^{-3} square. Trendlines with and without intercept are presented with dashed and solid black line, respectively

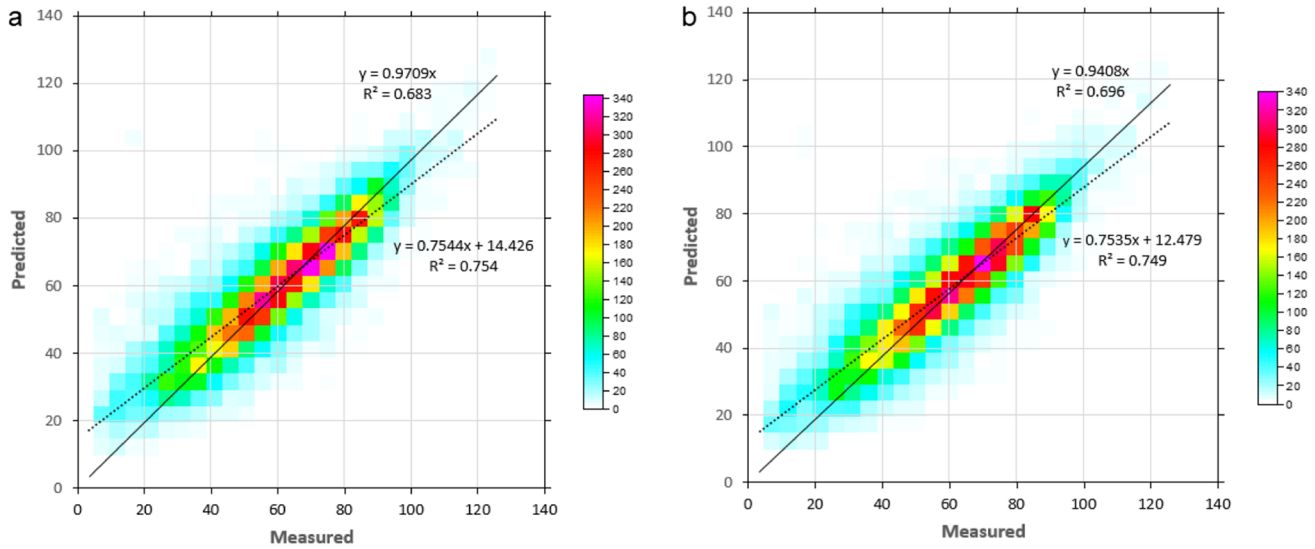

Fig. 2 Concentrations of ozone (μg m^{-3}) for years 2020–2021: **a** predicted from regression Eq. 2 and **b**) predicted from regression Eq. 1. Colour scale represents the number of cases in 5 by 5 μg m^{-3} square. Trendlines with and without intercept are presented with dashed and solid black line, respectively

Table 12 The parameters commonly used to evaluate the model performance (Willmott et al., 1985) applied to our models

Parameter	Eq. 1 to period 1	Eq. 1 to period 2	Eq. 2 to period 1	Eq. 2 to period 2
RMSE	11.84	10.96	12.2	10.63
RMSE_s	6.2068	5.6685	6.9870	5.2844
RMSE_u	10.2250	9.3789	10.2615	9.2620
d	0.9174	0.9224	0.9110	0.9262
R	0.8548	0.8656	0.8518	0.8686

mean square error, RMSE_d is unsystematic part of root mean square error, parameter *d* is index of agreement, and *R* is correlation coefficient. The columns 2–5 contain the parameters computed for particular cases, where "Eq. 1" means Eq. 1, "Eq. 2" means Eq. 2, and the equations were applied to the period 1 (2016–2019) and/or to the period 2 (2020–2021). We can see that Eqs. 1 and 2 are remarkably equipotent for the ozone concentration prediction purpose, but the prediction power depends on the period. For the period 2016–2019, that is longer and with more extensive concentration variations, all the residuals (characterized by parameters RMSE, RMSE_s, and RMSE_d) are larger and index of agreement (*d*) is smaller.

Ozone and NO_x participate in the same chemical reactions in the atmosphere, and therefore, a considerable anticorrelation between the time series of these variables was sometimes observed in the data of Tahkuse measurements, likewise the observed anticorrelation with CO_2 concentrations. Therefore, a remarkable regression between the ozone (O_3) and NO and/or NO_2 concentrations could be expected, but actually, the determination coefficient (R^2) was only up to 0.05. Also, considering the inclusion of the NO and/or NO_2 concentrations measured at Tahkuse to the list of independent variables of the above described multiple regression model (in addition to the concentrations of ozone measured at the other stations) only yield a negligible enhancement in the determination coefficient below about 0.05. The absence of remarkable link between NO and/or NO_2 concentrations and ozone concentrations is somewhat surprising, also because some studies (e.g., He and Lu 2012) have established notable relationship between the measured ozone and NO_2. The reasons that cause these different outcomes at Tahkuse station are not definitely known yet, but hypothetically the long-range transport of ozone (that has also experienced NO and/or NO_2 influence at that remote location) can override the ozone formation and sink pathways determined by nearby occurring NO and/or NO_2 concentrations.

The time series of developed ozone prediction model residuals (measured concentrations minus predicted by model concentrations) are shown in Figs. 3 and 4, and the results of more detailed analysis are presented in Figs. 5 and 6. Larger residuals in the first place belong to the predictions implemented at summer nights, often accompanied with small values of observed ozone concentrations at

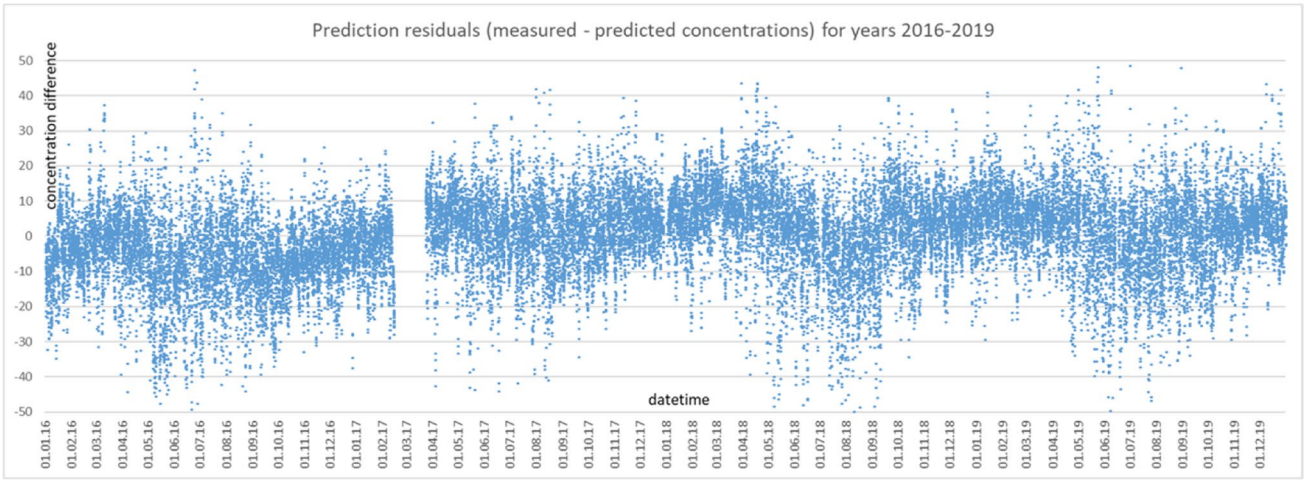

Fig. 3 The ozone prediction model residuals (μg m.$^{-3}$) for years 2016–2019

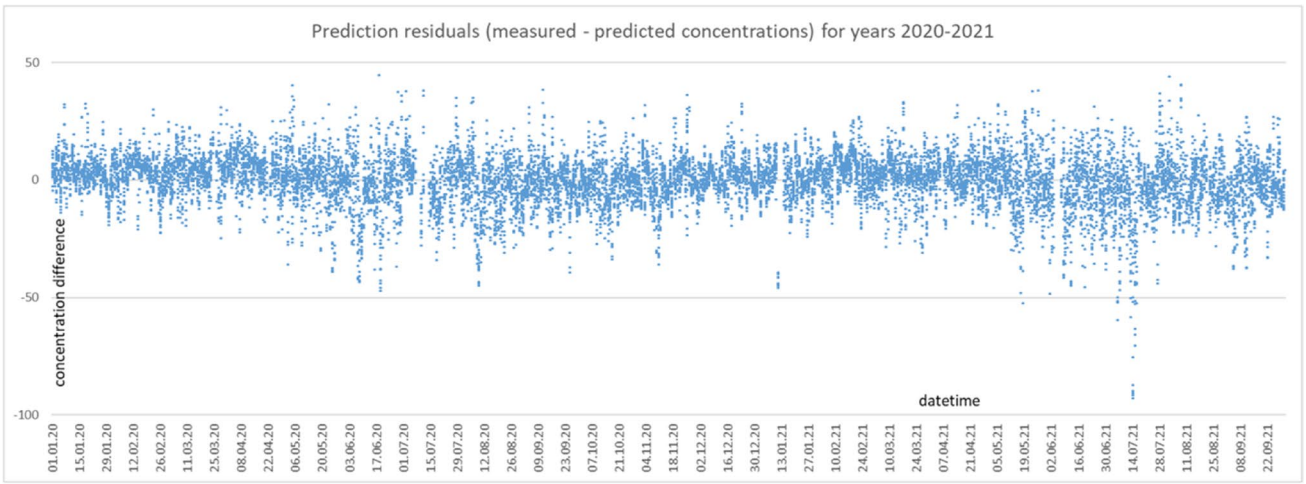

Fig. 4 The ozone prediction model residuals (μg m.$^{-3}$) for years 2020–2021

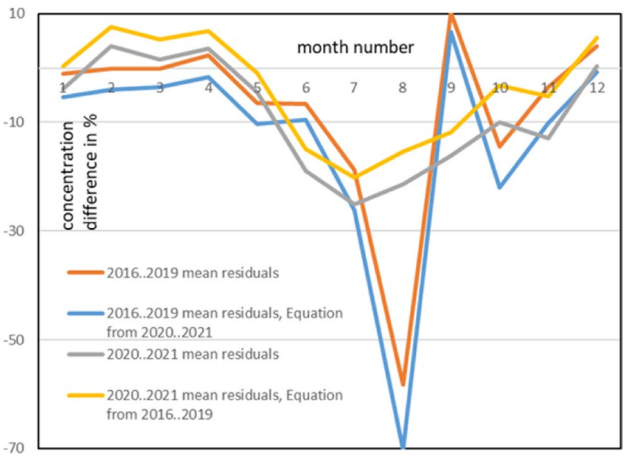

Fig. 5 The averaged extent of the residuals of the model as a function of month

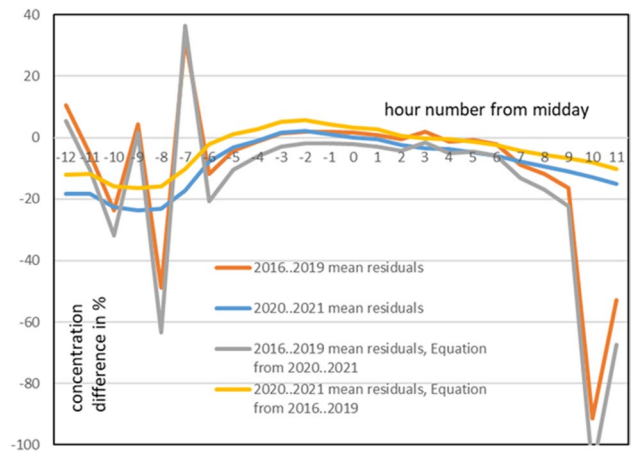

Fig. 6 The averaged extent of the residuals of the model as a function of time (hours from midday)

Tahkuse station. This combination results in large negative residuals, because in these cases, the model overestimates the values predicted for Tahkuse. Figure 6 demonstrates few abrupt changes in the residual values at certain specific hours, which can be due some extraordinary observed values within the period 2016–2019. In the period 2020–2021, all the curves that present variations in the residual values are far smoother.

Discussion and conclusions

To understand the differences in the factors that determine the concentrations of atmospheric trace and pollution gases and distribution patterns during the period 2016–2019 versus 2020–2021, mentioned several times above, we keep in mind two big changes, which took place:

(1) COVID-19 lockdown that started during first wave of pandemic in March–April 2020 and proceeded through temporary and partial relaxation and new restriction periods through 2020 and 2021;
(2) rapid increase of the price of European carbon emission quota, which caused dramatic decrease in oil-shale based energy production in Estonia, thus reducing industrial emissions from a certain area in North-Eastern Estonia a lot.

COVID-19 lockdown reduced the urban traffic flows and relevant NO_x emissions in urban centres worldwide, but simultaneously, it increased in some areas the residential heating due to bigger fraction of time spent at home; thus, the impact to the emissions of SO_2 and particulate matter emissions was more ambivalent (Sokhi et al. 2021). The concentrations of ozone even increased in some areas due to decrease in NO_2, an ingredient which is known as suppressing the ozone in urbanized areas.

The oil-shale power plants in the North-East Estonia are the dominant emission sources of sulphur dioxide and important sources of nitrogen oxides in Estonia (about 25% from all NOx pollution). The major source of NO_x is transport (about 40%). According to national statistical survey, emissions started dramatically decrease in 2019, when about 60% less of SO_2 was emitted from power plants and 45% from the whole Estonia, when compared to the average of 2016–2018. In 2020–2021, the emissions decreased even more, as these power plants were nearly suspended most of time.

In five-factor set determined by factor analysis for NO_2 concentrations, the first and second factors are swapped in 2020–2021 set, compared to 2016–2019. Obviously, the first factor that describes the earlier period (similar to the second factor that describes the later period) represents the influence of street transport (dominated by Liivalaia and Õismäe urban stations in capital city Tallinn). The second factor is dominated by continental rural background stations Saarejärve, Tahkuse, and Lahemaa, whereas industrial Kohtla-Järve and maritime Vilsandi indicate weakly opposite effects. Swap of these two factors may have occurred due to cleaning the urban air due to lockdowns, which diminishes the importance of urban-type pollution patterns. The third factor of both periods is dominated by rather unique maritime Vilsandi site, whereas fourth factor might be interpreted as influenced by residential heating (Tartu), opposing to the cleanest continental site, by means of local pollution sources, in the Lahemaa national park. The fifth factor is likely dominated by industrial emissions from oil-shale-based industries and power plants (Kohtla-Järve), with slightly opposing street station Liivalaia and rural background at Tahkuse, which is at largest distance from oil-shale processing facilities, considering the continental stations. Nevertheless, we have to consider that the description ability of the last factors is quite weak, and therefore, the uncertainties are considerable.

The factors of NO are partially similar to NO_2, with a few exceptions. There are obvious reasons for different time-dependent patterns in different site types; e.g., the diurnal course is more pronounced in urban sites due to changes in traffic and residential heating. The first and the second factors, representing respectively urban and rural-continental influences, are not swapped between periods, but stay respectively at first and second place. We have to keep in mind that in contrary to the oxidation product NO_2, which is also dependent on long-range transport of air pollutants, the NO represents a "fresh" pollution that dominates near the source. It is reasonable to assume that close to the urban street emission sources, the primary emissions retain their importance despite somewhat lower concentrations. On the other hand, the Kohtla-Järve site, which nearly entirely determines the fifth factor in later period (2020–2021), contributes more to the "urban" factor 1 in earlier period (2016–2019). It can be that after the lockdown-induced decrease of street transport, the industrial emissions started to dominate in that industrial town even above the usually street-emission-dominated NO concentrations. Finally, the Lahemaa-dominated "clean-continental" factor moved from fifth to fourth place, which may be due to cleaning the air due to lockdown.

First of the SO_2 factors is defined nearly equally by urban stations Liivalaia and Õismäe in Tallinn — as a rule, the concentrations of SO_2 are rather low there, but additionally slight influence of urban-scale industrial and residential heating sources takes place. Next factors are different in mentioned two time periods, which may be due to dramatic decrease of SO_2 emissions from oil-shale based power plants at about 2020. These power plants are located close to Kohtla-Järve, about 110 km away from Lahemaa and Saarejärve background

stations, but far away from Tahkuse (about 200 km) and Vilsandi (about 375 km) rural stations. Indeed, these two far-away stations dominate in the second factor during 2016–2019, but get separated in 2020–2021 by factors 3 and 5, as most probably the other pollution mechanisms start to dominate over weakened impact of power plant emissions. Remarkably, the "industrial" factor dominated by Kohtla-Järve, is moved from third (2016–2019) to fourth (2020–2021) place, which likely refers to decreased influence of those emissions. It should be kept in mind that Kohtla-Järve is located in the region of oil-shale-based industries, which emissions, SO_2 most remarkable among them, were a lot lower in second period (2020–2021). Lahemaa is the closest rural station to the oil-shale operated power plants (60–110 km), often remarkably affected by their emissions, and sometimes also influenced by shipping emissions from Gulf of Finland.

Ozone is generated by chemical reactions initiated by solar UV radiation; therefore, the main factor could be the rural background conditioned by the definite meteorological situation mixed by urban-industrial conditions. The first, most powerful factor with high certainty corresponds to the meteorological situation in the presence of anticyclonic air system that covers the entire Estonia and, therefore, influence the change (rise) of O_3 concentration in all the stations. Long-range transport of ozone can also belong to this factor. The factors 2 and 3 are largely determined by Vilsandi and Liivalaia, the first station represents clear maritime conditions, and the second one represents the strong urban (traffic) effects. These effects have rather opposite character, which can bring up the opposite signs within the factor 3. Uniqueness of Liivalaia station, which is strongly related with factors 2 and 3, is obviously related to its location next to a street with heavy vehicular emissions. In such sites, the production of ozone is hindered by high concentration of NO. Contribution of Vilsandi to factor 2 is not well understood and needs further research. Factor 4 patterns are rather similar for both periods with high scores for Tartu and Tahkuse with opposite signs. This may be due to the absence of high traffic on streets and a lot of residential heating just nearby the measurement station, whereas the opposite signs of the participations may be caused by the extents of the characteristic effects. The factor 5 could correspond to some other (industrial) pollution, because the values of the corresponding factor scores are large for the industrial region station Kohtla-Järve. The effects of Kohtla-Järve seem to be opposite to the ones characteristic for Tahkuse (period 2016–2019) and for both Tahkuse and Tartu (period 2020–2021).

The model designed for estimation and forecast of ozone values at a specific location (Tahkuse), based on the known concentrations measured at several other locations (in this case at Saarejärve, Vilsandi and Lahemaa), was able to predict the general trends of ozone concentrations (the determination coefficients were $R^2 = 0.745$ for the period 2016–2019 and 0.762 for the period 2020–2021). However, the model is still is not able to predict several specific concentrations that are apparently driven by certain local factors. The latter especially applies to the cases (time intervals) when the ozone concentration measured at Tahkuse was low, but the concentrations measured at other locations were not that low. These cases are characterized by large negative residuals as is observable in the Figs. 3 and 4. These large negative residuals, first of all, tend to occur at summer nights with high CO_2 levels (Figs. 5 and 6). At daytime (from about midday minus 6 h to about midday plus 7 h) and during the colder season, the regression model performs significantly better. The cases when the concentrations measured at Tahkuse exceed the predicted values take place as well, but the cases with large negative residuals are far more prominent. Therefore, even though the variations in ozone concentrations are predominantly determined by only one factor (the first one) as discussed above, the local factors cannot be omitted, and therefore, the actual continuous ozone monitoring cannot be substituted by estimates based on the results observed at other locations. The latter is certainly valid for the Tahkuse station, even though generally the values can be estimated. The estimates can be used at certain extraordinary cases, e.g., when the data are not available because of some break in continuous measurements. Here, we performed the regression analysis of Tahkuse ozone data, but it is also interesting to know, in what extent the concentration trends at all nearby stations are linked (and can be predicted from other data), but this study is to be accomplished in future.

Acknowledgements The authors are thankful to Hilja Iher for kind assistance in maintaining the equipment in Tahkuse monitoring station. The measurements at Tahkuse station are supported by the Estonian Environment Agency contract LLTFY21274 "Complex studies of the air quality at Tahkuse in 2021" contributing to the fulfilment of National Environmental Monitoring Program.

Funding This work was supported by the European Regional Development Fund (project MOBTT42) under the Mobilitas Pluss programme, Estonian Research Council (project PRG714), Estonian Environmental Observatory (KKOBS, project 2014–2020.4.01.20–0281). Also, the Estonian-Swiss Cooperation Programme "Enhancing public environmental monitoring capacities" project MLOFY12171 "Updating of the equipment of the Tahkuse Air Monitoring Station with the monitor of polluting gases" (2012–2016) financially supported this work.

Data availability Data will be made available on request and/or included as electronic supplement according to the data policy of the journal.

Declarations

Ethical approval This study needs no special ethical approval.

Consent to participate All the authors have declared their consent to participate in this study.

Consent for publication All the authors have declared their consent to publish this paper.

Conflict of interest The authors declare no competing interests.

References

Araki S, Iwahashi K, Shimadera H, Yamamoto K, Kondo A (2015) Optimization of air monitoring networks using chemical transport model and search algorithm. Atmos Environ 122:22–30. https://doi.org/10.1016/j.atmosenv.2015.09.030

Chan TW, Mozurkewich M (2007) Application of absolute principal component analysis to size distribution data: identification of particle origins. Atmos Chem Phys 7:887–897 www.atmos-chem-phys.net/7/887/2007/

He HD, Lu WZ (2012) Decomposition of pollution contributors to urban ozone levels concerning regional and local scales. Build Environ 49:97–103. https://doi.org/10.1016/j.buildenv.2011.09.019

Junninen H, Niska H, Tuppurainen K, Ruuskanen J, Kolehmainen M (2004) Methods for imputation of missing values in air quality data sets. Atmos Environ 38:2895–2907. https://doi.org/10.1016/j.atmosenv.2004.02.026

Li Q, Li X, Li H (2022) Factors Influencing PM2.5 Concentrations in the Beijing–Tianjin–Hebei urban agglomeration using a geographical and temporal weighted regression model. Atmosphere 13:407. https://doi.org/10.3390/atmos13030407

Liu B, Sun X, Zhang J, Bi X, Li Y, Li L, Dong H, Xiao Z, Zhang Y, Feng Y (2020) Characterization and spatial source apportionments of ambient PM10 and PM2.5 during the heating period in Tian'jin. China Aerosol and Air Quality Research 20:1–13. https://doi.org/10.4209/aaqr.2019.06.0281

Nogarotto DC, Pozza SA (2020) A review of multivariate analysis: is there a relationship between airborne particulate matter and meteorological variables? Environ Monit Assess 192:57. https://doi.org/10.1007/s10661-020-08538-1

Sharma E, Deo RC, Prasad R, Parisi AV et al (2020) A hybrid air quality early-warning framework: an hourly forecasting model with online sequential extreme learning machines and empirical mode decomposition algorithms. Sci Total Environ 709:135934. https://doi.org/10.1016/j.scitotenv.2019.135934

Shubhankar B, Ambade B (2016) Spatio-temporal variability of ambient trace gas pollutants and their PCA predication: a comprehensive review. Rasayan J Chem 9:112–120 (http://www.rasayanjournal.com)

Sokhi R, Singh V, Querol S et al (2021) A global observational analysis to understand changes in air quality during exceptionally low anthropogenic emission conditions. Environ Int 157:106818. https://doi.org/10.1016/j.envint.2021.106818

Tai-Yi Yu (2010) Characterization of ambient PM2.5 concentrations. Atmos Environ 44:2902–2912. https://doi.org/10.1016/j.atmosenv.2010.04.034/

Wang J, Du P, Hao Y, Ma X, Niu T, Yang W (2020) An innovative hybrid model based on outlier detection and correction algorithm and heuristic intelligent optimization algorithm for daily air quality index forecasting. J Environ Manag 255:109855. https://doi.org/10.1016/j.jenvman.2019.109855

Wei-Zhen Lu, Hong-Di He, Li-yun D (2011) Performance assessment of air quality monitoring networks using principal component analysis and cluster analysis. Build Environ 46:577–583. https://doi.org/10.1016/j.buildenv.2010.09.004

Willmott CJ, Ackleson SG, Davis RE, Feddema JJ, Klink KM, Legates DR, O'Donnel J, Rowe CM (1985) Statistics for the evaluation and comparison of models. J Geophys Res 90:8995–9005

Xu J, Yang W, Bai Z, Zhang R, Zheng J, Wang M, Zhu T (2022) Modeling spatial variation of gaseous air pollutants and particulate matters in a metropolitan area using mobile monitoring data. Environ Res 210:112858. https://doi.org/10.1016/j.envres.2022.112858

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.