Using General Least Deviations Method for Forecasting of Crops Yields

Tatiana Makarovskikh, Mostafa Abotaleb,
Anatoly Panyukov

Department of Computer Science,
National Research University SUSU
Chelyabinsk, Russian Federation
[Makarovskikh.t.a,abotalebm,paniukovav]@susu.ru

22nd INTERNATIONAL CONFERENCE
MATHEMATICAL OPTIMIZATION THEORY AND OPERATIONS RESEARCH
(MOTOR 2023)
Ekaterinburg, Russia, 03 – 07 July, 2023.

Using General Least Deviations Method for Forecasting of Crops Yields

Content

Introduction: on Crop Yields

Review

- Ahmad, R., Yang, B., Ettlin, G., Berger, A., Rodriguez-Bocca, P. A machine learning based ConvLSTM architecture for NDVI forecasting (2020) https://doi.org/10.1111/itor.12887.
- Gao, P., Du, W., Lei, Q., Li, J., Zhang, Sh., Li, N. NDVI Forecasting Model Based on the Combination of Time Series Decomposition and CNN-LSTM (2023) https://doi.org/10.1007/s11269-022-03419-3.
- Ahmad, R., Yang, B., Rodriguez-Bocca, P. Deep Spatial-Temporal Graph Modeling for Effcient NDVI Forecasting (2023) https://doi.org/10.1016/j.atech.2023.100172.
- Huang, Sh., Ming, Bo, Huang, Q., Leng, G., Hou, B. A Case Study on a Combination NDVI Forecasting Model Based on the Entropy Weight Method (2017). https://doi.org/10.1007/s11269-017-1692-8.
- Fernandez-manso, A., Quintano, C., Fernandez-Manso, O. Forecast of NDVI in coniferous areas using temporal ARIMA analysis and climatic data at a regional scale (2011) https://doi.org/10.1080/01431160903586765.
- Alhamad, M., Stuth, J., Vannucci, M. Biophysical modelling and NDVI time series to project near-term forage supply (2007) https://doi.org/10.1080/01431160600954670.

Introduction: on Crop Yields

Review. Russian publications

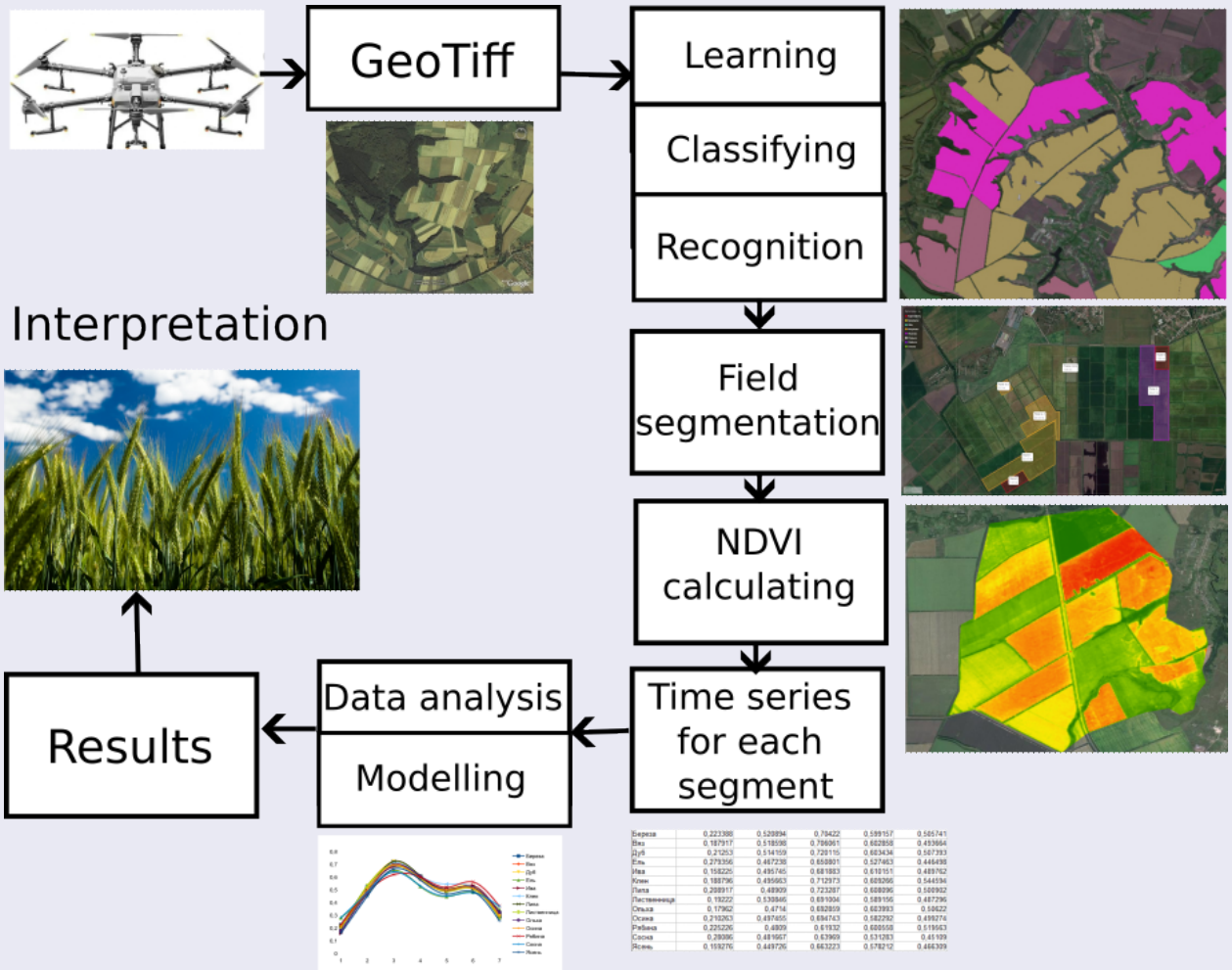
- Bukhovets, A. G., Semin, E.A., Kostenko, E.I., Yablonovskaya, S.I. Modelling of the dynamics of the NDVI vegetation index of winter wheat under the conditions of the CFD (2018) https://doi.org/10.17238/issn2071-2243.2018.2.186 (in Russian).
- Greben, A.S., Krasovskaya, I.G. Analysis of the main methods for forecasting yields using space monitoring data, in relation to grain crops in the steppe zone of Ukraine. Radio electronic and computer systems. 2 (54). 170–180 (2012). (in Russian).
- Spivak, L.F., Vitkovskaya, I.S., Batyrbayeva, M.Zh., Kauazov, A.M. Analysis of the results of forecasting the yield of spring wheat based on time series of statistical data and integral indices of vegetation. Modern problems of remote sensing of the Earth from space. 12 (2). 173–182 (2015).(in Russian).

Monitoring of crop yields

- a higher level of detail, which is of particular interest to potential customers
- detecting the problematic areas of the field
- methods for identifying the parameters of a single quasilinear difference equation
- allows to get the model coefficients for any considered objects

Data organizing

The scheme of the analysis system executing



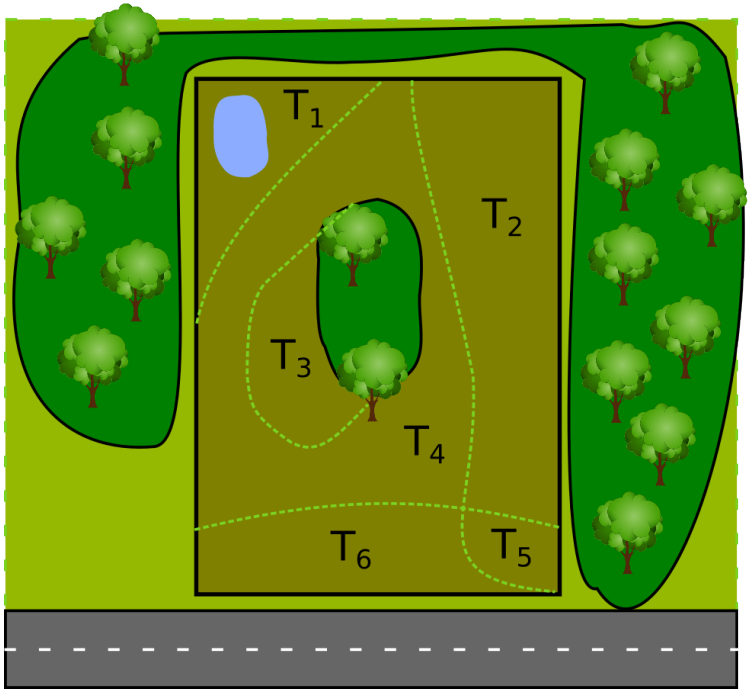
Normalized Difference Vegetation Index (NDVI)

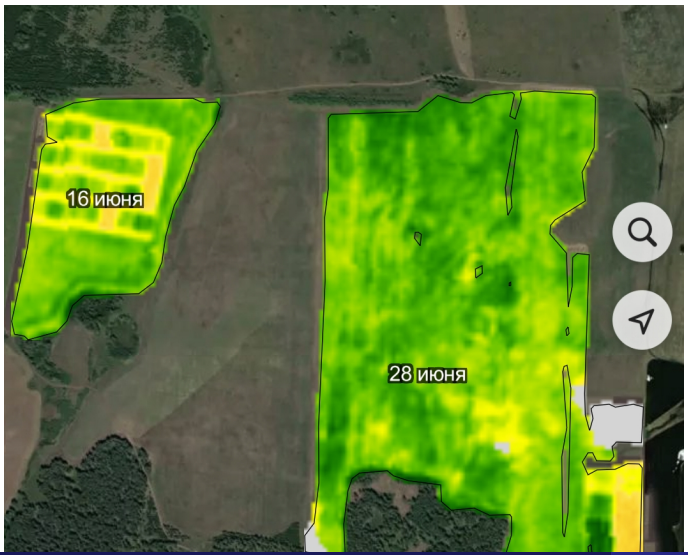
$$NDVI = \frac{NIR - RED}{NIR + RED}$$

Properties:

- $NDVI \in [-1; 1]$
- $NDVI \leq 0$: buildings, structures, paved road surfaces, water surfaces, mountains, clouds and snow.
- $NDVI \in [0.1; 0.2)$: an open soil
- $NDVI \in [0.2; 0.4)$: the weak, sparse vegetation,
- $NDVI \in [0.4; 0.6)$: moderate vegetation,
- $NDVI \geq 0.6$: healthy, dense vegetation.

Field Clustering

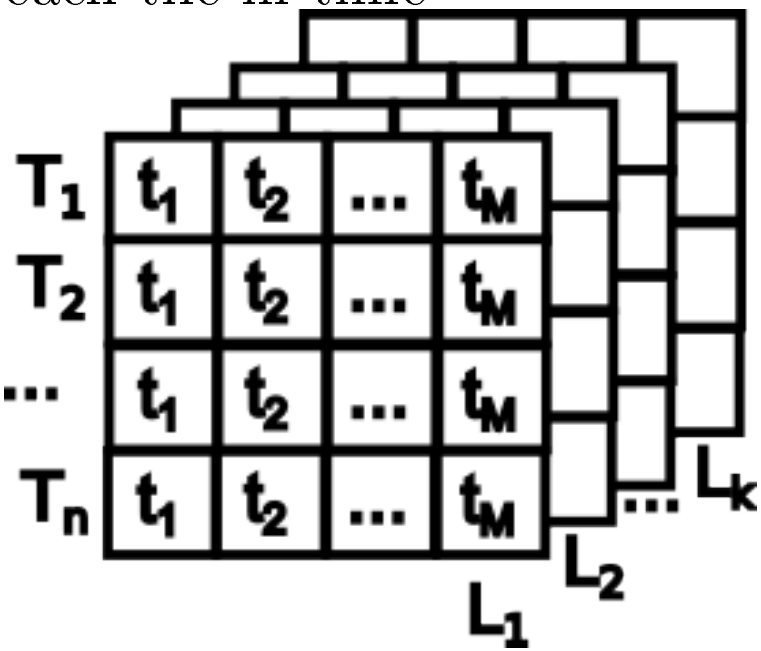




- T_1 corresponds the area near the pond, it's most likely low;
- T_2 and T_3 satisfy the shadowed area;
- T_4 is the common field without any peculiarities;
- T_5 is the shadowed area near the road;
- T_6 is the area near the road.

Data Organizing

The matrix of values for each tile in time



- Monitoring of crops developing is made in 3–5 days
- the number of images M is less than 50 per season
- the number of tiles N depends on:
  - the real size of the recognized objects
  - properties of each object
- to describe the dynamic process for each tile of the recognized object we need to obtain the coefficients for N models
- the task is solvable for each agricultural object separately

General least deviations method estimation

Statement of the problem

To determine the coefficients $a_1, a_2, a_3 \ldots, a_m \in \mathbb{R}$ of a m-th order quasilinear autoregressive model

$$y_t = \sum_{j=1}^{n(m)} a_j g_j(\{y_{t-k}\}_{k=1}^{m}) + \varepsilon_t, \quad t = 1, 2, \ldots, T$$

by up-to-date information about of values of state variables $\{y_t \in \mathbb{R}\}_{t=1-m}^{T}$ at time instants t; here $g_j : (\{y_{t-k}\}_{k=1}^{m}) \to \mathbb{R}$, j=1,2,...n(m) are given $n(m)$ functions, and $\{\varepsilon_t \in \mathbb{R}\}_{t=1}^{T}$ are unknown errors.

General least deviations method

Approach

Input: time series $\{y_t \in \mathbb{R}\}_{t=-1-m}^{T}$ of length $T + m \geq \left(1 + 3m + m^2\right)$
Output: factors $a_1, a_2, a_3 \ldots, a_m \in \mathbb{R}$
Optimization task

$$\sum_{t=1}^{T} \arctan \left| \sum_{j=1}^{n(m)} a_j g_j(\{y_{t-k}\}_{k=1}^{m}) - y_t \right| \to \min_{\{a_j\}_{j-1}^{n(m)} \subset \mathbb{R}}$$

The Cauchy distribution

$$F(\xi) = \frac{1}{\pi} \arctan(\xi) + \frac{1}{2}$$

has the maximum entropy among distributions of random variables that have no mathematical expectation and variance.

General least deviations method

The basic set $g_j(*)$

$$g_{(k)}(\{y_{t-k}\}_{k=1}^{m}) = y_{t-k},$$

$$g_{(kl)}(\{y_{t-k}\}_{k=1}^{m}) = y_{t-k} \cdot y_{t-l},$$

$$k = 1, 2, \ldots, m;\ \ l = k, k+1, \ldots, m.$$

- $n(m) = 2m + \mathrm{C}_m^2 = m(m+3)/2$
- the numbering of $g_{(*)}$ functions can be arbitrary

For $m = 2$ we have the following functions $g_{(*)}$:

$$g_1 = y_1, \quad g_2 = y_2, \quad g_3 = y_1^2, \quad g_4 = y_2^2, \quad g_5 = y_1 \cdot y_2.$$

GLDM estimation task

- concave optimization problem
- entering the additional variables reduces it to LP task

$$\sum_{t=1}^{T} p_t z_t \to \min_{\substack{(a_1,a_2,\ldots,a_{n(m)})\in \mathrm{R}^m,\\ (z_1,z_2,\ldots,z_T)\in \mathrm{R}^T}}$$

$$-z_t \le \sum_{j=1}^{n(m)} [a_j g_j(\{y_{t-k}\}_{k=1}^{m})] - y_t \le z_t, \quad t = 1, 2, \ldots, T,$$

$$z_t \ge 0, \quad t = 1, 2, \ldots, T.$$

This task has a canonical type with variables $n(m) + T$ and $3n$ inequality constraints including the conditions of non-negativity of z_j, $j = 1, 2, \ldots, T$.

The dual task

$$\sum_{t=1}^{T} (u_t - v_t)\, y_t \to \max_{u,v \in \mathbb{R}^T},$$

$$\sum_{t=1}^{T} a_j g_j(\{y_{t-k}\}_{k=1}^{m})\,(u_t - v_t) = 0,\ j = 1, 2, \ldots, n(m),$$

$$u_t + v_t = p_t, \quad u_t,\, v_t \geq 0, \quad t = 1, 2, \ldots, T.$$

$$w_t = u_t - v_t,\ t = 1, 2, \ldots, T.$$

$$u_t = \frac{p_t + w_t}{2}, \quad v_t = \frac{p_t - w_t}{2}, \quad -p_t \leq w_t \leq p_t, \quad t = 1, 2, \ldots, T.$$

So the optimal solution of primal task is equal to the optimal solution of:

$$\sum_{t=1}^{T} w_t \cdot y_t \to \max_{w \in \mathbb{R}^T},$$

$$(1): \quad \sum_{t=1}^{T} g_j(\{y_{t-k}\}_{k=1}^{m}) \cdot w_t = 0,\ j = 1, 2, \ldots, n(m),$$

$$(2): \quad -p_t \leq w_t \leq p_t,\ t = 1, 2, \ldots, T.$$

Constraints

Constraints (1) define $(T - n(m))$-dimensional linear variety $\mathcal{L}$ with $(n(m) \times T)$-matrix

$$S = \begin{bmatrix} g_1(\{y_{1-k}\}_{k=1}^m) & g_1(\{y_{2-k}\}_{k=1}^m) & \cdots & g_1(\{y_{T+1-k}\}_{k=1}^m) \\ g_2(\{y_{1-k}\}_{k=1}^m) & g_2(\{y_{2-k}\}_{k=1}^m) & \cdots & g_2(\{y_{T+1-k}\}_{k=1}^m) \\ \vdots & \vdots & \ddots & \vdots \\ g_{n(m)}(\{y_{1-k}\}_{k=1}^m) & g_{n(m)}(\{y_{1-k}\}_{k=1}^m) & \cdots & g_{n(m)}(\{y_{1-k}\}_{k=1}^m) \end{bmatrix}$$

Constraints (2) define T-dimensional parallelepiped $\mathcal{T}$.

Solution

We obtain the solution by algorithm using the gradient projection of the objective function

$$\nabla = \{y_t\}_{t=1}^{T}$$

on the allowed area $\mathcal{L} \cap \mathcal{T}$ defined by the constraints (1)–(2).
The projection matrix:

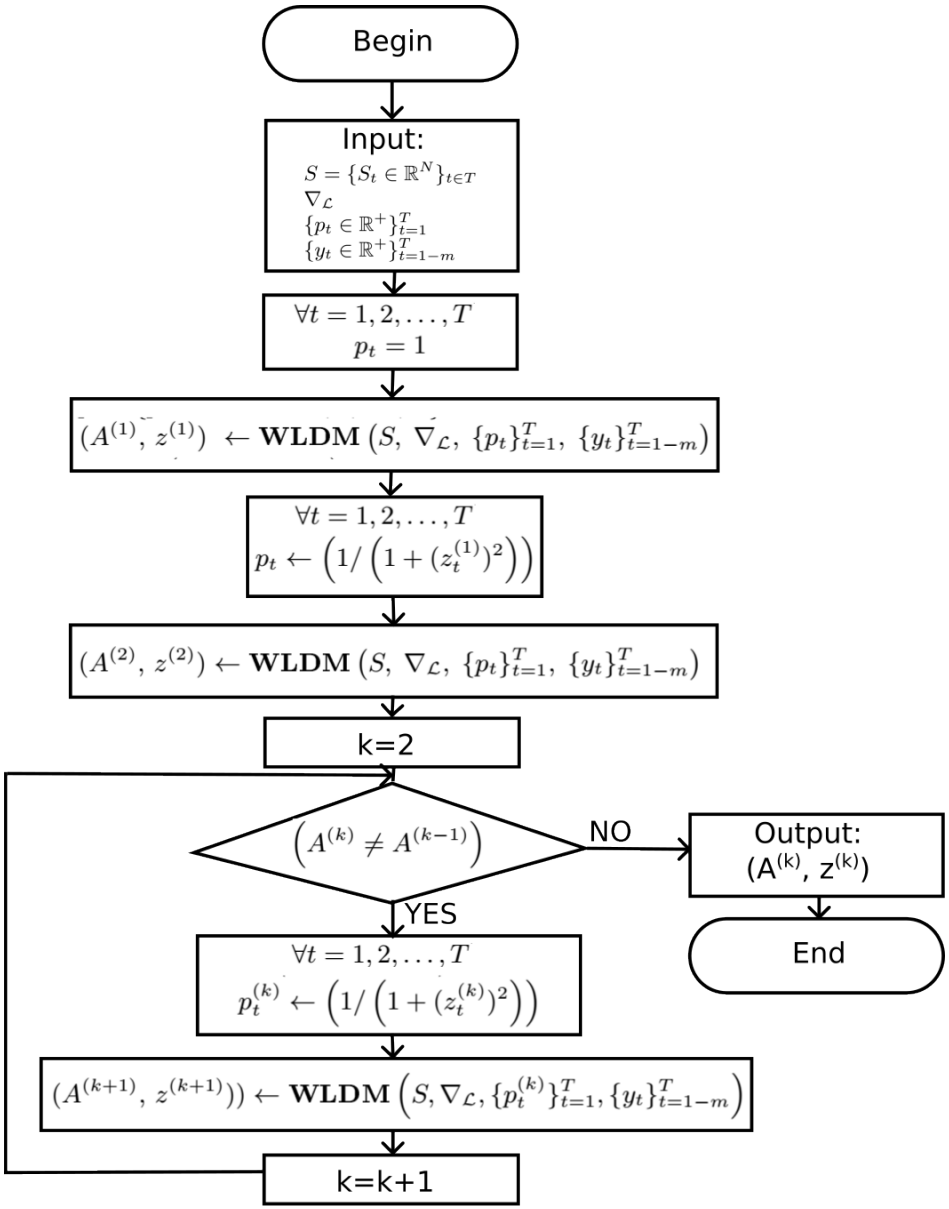
$$S_{\mathcal{L}} = E - S^T \cdot \left(S \cdot S^T\right)^{-1} \cdot S,$$

and gradient projection on $\mathcal{L}$ is:

$$\nabla_{\mathcal{L}} = S_{\mathcal{L}} \cdot \nabla$$

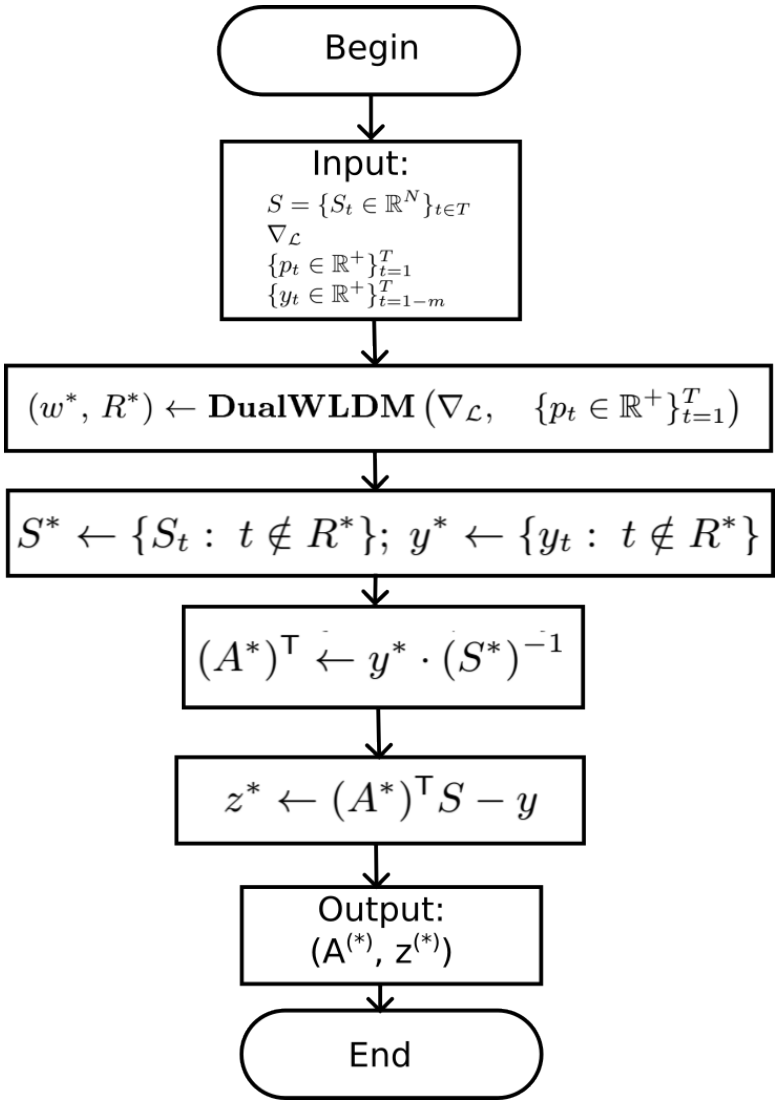
If outer normal on any parallelepiped face forms the sharp corner with gradient projection $\nabla_{\mathcal{L}}$ then movement by this face is equal to zero.

The scheme of GLDM estimation algorithm



- Algorithm runs as the iteration process for obtaining optimal GLDM solution $A \in \mathbb{R}^{n(m)}$ and the vector of residuals $z \in \mathbb{R}^T$. This process stops when $\left(A^{(k)} = A^{(k-1)}\right)$.
- To obtain A and z we run the WLDM estimation algorithm

WLDM estimation algorithm



- calculates the factors

$$a_1, a_2, a_3 \ldots, a_{n(m)} \in \mathbb{R}$$

by solving the optimization task

$$\sum_{t=1}^{T} p_t \cdot \left| \sum_{j=1}^{n(m)} a_j g_j(\{y_{t-k}\}_{k=1}^{m}) - y_t \right|$$

$$\rightarrow \min_{\{a_j\}_{j=1}^{n(m)} \in \mathbb{R}^{n(m)}}$$

Solution

- Let (w^*, R^*) be the result of executing the gradient projection algorithm
- w^* be the optimal solution to the dual task
- the optimal solution of the primal task is

$$u_t^* = \frac{p_t + w_t^*}{2}, \quad v_t^* = \frac{p_t - w_t^*}{2}, \quad t = 1, 2, \ldots, T.$$

- $(\{a_j^*\}_{j-1}^{n(m)}, z^*)$ the optimal solution of the task WLDM

The system of linear algebraic equations

It is following from the complementarity condition for a pair of mutually dual tasks that

$$y_t = \sum_{j=1}^{n(m)} [a_j g_j(\{y_{t-k}\}_{k=1}^{m})] \qquad \forall t \notin R^*,$$

$$y_t = \sum_{j=1}^{n(m)} [a_j g_j(\{y_{t-k}\}_{k=1}^{m})] + z_t^*, \quad \forall t \in R^* : w_t^* = p_t,$$

$$y_t = \sum_{j=1}^{n(m)} [a_j g_j(\{y_{t-k}\}_{k=1}^{m})] - z_t^*, \quad \forall t \in R^* : w_t^* = -p_t.$$

Solution

Theorem 1

Let

- w^* be the optimal solution of the dual task,
- $(\{a_j^*\}_{j-1}^{n(m)}, z^*)$ be solution of a system of linear algebraic equations.

Then $\{a_j^*\}_{j-1}^{n(m)}$ is the optimal solution to the task WLDM.

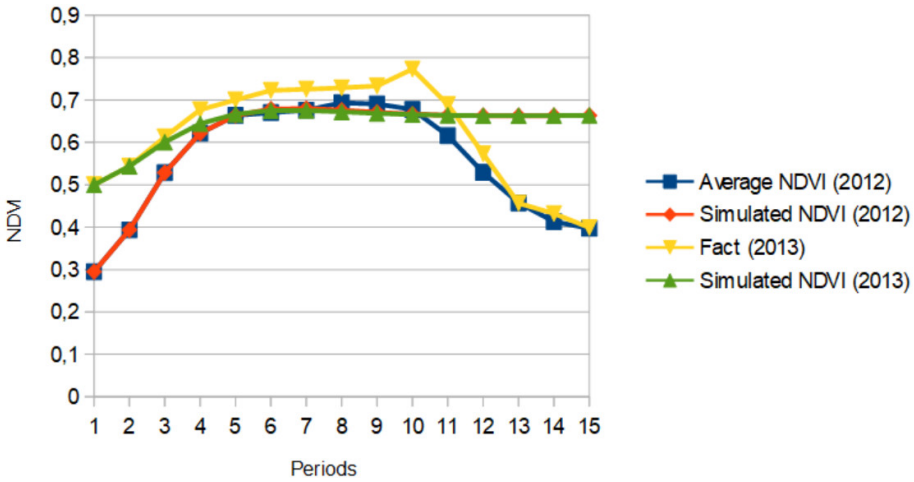
Theorem 2

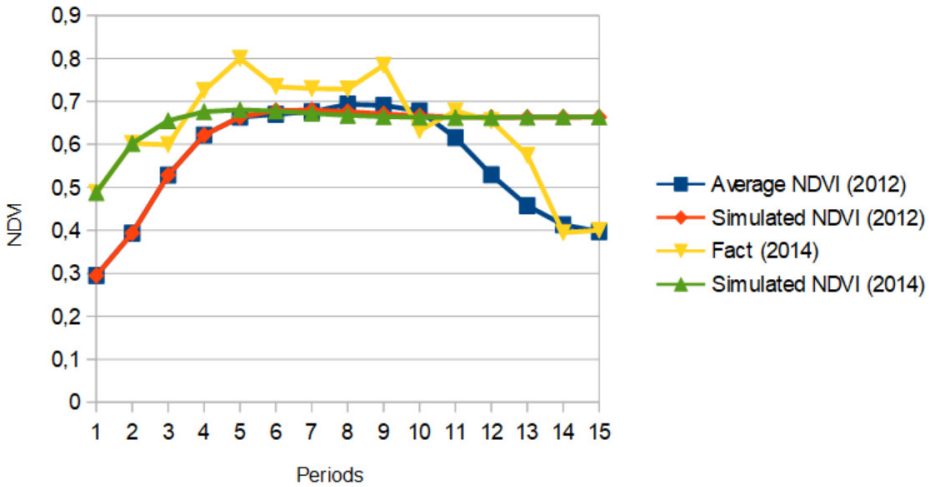
The sequence$\{\left(A^{(k)}, z^{(k)}\right)\}_{k=1}^{\infty}$, constructed by GLDM-estimator Algorithm, converges to the global minimum (a^*, z^*) of the task GLDM.

Computational experiment

- the dynamics of index for winter wheat sowing in Stavropol region
- the dynamics of NDVI for forests of Losiniy island

The dynamics of index for winter wheat sowing in Stavropol region





$$y_t = (3.46935 \cdot y_{t-1} - 2.18641 \cdot y_{t-2}) -$$

$$-5.59237 \cdot y_{t-1}^2 - 2.5635 \cdot y_{t-1} y_{t-2} +$$

$$+7.72991 \cdot y_{t-2}^2$$

Significance levels:

- 2012: 0,9445357262
- 2013: 0,7603522142
- 2014: 0,6913386433

The dynamics of NDVI for modelling the growth of trees in Losiniy island

Tree	a_1	a_2	a_3	a_4	a_5	MAE	MBE
Birch	-3,33289	4,3374	13,6638	8,41678	-22,82575	1,00E-012	-1,00E-012
Elm	-2,72226	3,50835	11,0937	6,64245	-18,05596	1,42E-014	1,42E-014
Oak	-3,04254	4,35999	13,2885	8,65894	-23,30116	1,70E-014	-1,70E-014
Spruce	-0,55497	3,03523	12,3839	10,3757	-26,43966	4,13E-014	-3,51E-014
Willow	-2,21953	2,71921	9,97028	6,33381	-16,02432	3,93E-014	3,93E-014
Maple	-19,3533	22,2068	54,9690	30,3626	-90,5544	1,71164	-1,71164
Linden	-0,88513	2,09379	10,0772	9,18861	-20,48208	7,26E-013	7,26E-013
Larch	-4,45510	5,07474	13,9688	6,54746	-20,55857	3,65776	-3,65776
Alder	-0,94055	1,94152	9,30427	8,30144	-18,28108	1,21342	1,21342
Aspen	-4,03793	5,60943	16,2789	10,01083	-28,17523	3,07E-013	2,53E-013
Rowan	-4,62792	4,21466	19,5855	12,74890	-30,28771	6,22E-015	6,22E-015
Pine	-0,96789	2,93664	11,2552	8,35734	-22,09748	7,66E-015	-7,66E-015
Ash	-0,91149	1,48240	9,41523	8,47214	-17,89959	2,16E-014	2,16E-014

Conclusion

- The model can be used to approximate the missing values of the NDVI, estimate the time to reach the maximum value of the index and, therefore, predict the start of harvesting dates.
- its errors are not worse than ones for neural network approaches or classical statistical models but needs less computational resources
- a significant advantage in comparison with these models that is the opportunity to interpret the model coefficients in terms of the research problem

Further research:

- improving the algorithm for time series using arbitrary time periods
- application of the developed approach for multidimensional time series
- the research of different outer factors such as humidity and temperature influence