DOI: 10.1478/AAPP.912A3

AAPP | Atti della Accademia Peloritana dei Pericolanti
Classe di Scienze Fisiche, Matematiche e Naturali
ISSN 1825-1242

Vol. 91, No. 2, A3 (2013)

A CATEGORICAL-INFORMATIONAL APPROACH TO THE VALUE PREDICTION PROBLEM

Gennadii V. Kondratiev *

(communicated by Gaetana Restuccia)

Abstract. A non-statistical approach to the problem of estimation of unknown parameters of empirical data is given. It is based on an invariant geometric information associated to the data.

1. Introduction

The purpose of this paper is twofold. On one hand it aims to deal with a problem in applied mathematics such as a price evolution prediction. On the other hand it wishes to bring to the fore examples of how techniques borrowed from the so called pure mathematics can be very concretely used for applications. The paper will focus on the problem of extending data from the real estate market, modeled by a metric or multimetric space. The task is to be able to model a situation starting from a few empirical data which are supposed to have been measured.

The proposed method is based on the relationship 'sample $\subset$ population'. It is supposed that this inclusion is *internal* with respect to a class of admissible transformations/*infomorphisms* (the new Definition 2, see below) which do not change the information contained in the data. Internal here means invariant modulo the given sample and the class of admissible transformations, so that the inclusion can be approximately reconstructed as a 'sample $\subset$ approximate population $\approx$ population'. The admissible transformations depend on the problem and constitute a priory information. There are examples of them in the text, but they are not formally defined, and taken in an abstract way as generating a suitable subcategory of the data model category.

The original contribution of the paper is in that the above mentioned inclusion 'sample $\subset$ approximate population $\approx$ population' is *natural* with respect to the class of infomorphisms (the new Definition 2, see below). It gives a simple and powerful way to build it. The method also allows to directly work with the data density in a certain and unique way without priory statistical assumptions. The simpler and the subtler the model constraints are, smaller the error of prediction is. When in statistics the error inside of the model is estimated by the variance, it means

only that if the real population is fully described by this model, then the error is distributed like that. In the real life case, the error is much bigger. This is because of the model error. For example, the widely used Gaussian assumption on the noise is wrong if there is an intentional influence to the population. In the proposed method, the original assumptions are as subtle as possible. The only supposition is that the inclusion 'sample $\subset$ population' is internal in the above sense. Although it sounds quite metaphysical, the choice of this principle brings a better 'uniform' estimation than that obtained by the usual technique.

Two different sets of *data*, thought of as subsets of a (multi)metric space, are considered the same if they are 'visually' (geometrically) identical. It is then natural to study a *category* whose *morphisms* model the set of admissible transformations of the base spaces, in such a way that the transformed data are considered identical to the original ones. The extension of data is then achieved by means of fiberwise natural transformations on a fibered category of subsets in multimetric spaces considered together with a fixed partition/foliation. Such a data extension will be thought of as a pointwise approximation, invariant under admissible transformations. We recall in Section **2** a few basic ready to use definitions in category theory, with an eye to applications. They are intended for non-specialists.

In Section **3** the statement of the problem and a categorical model of the data are given. In Section **4** a lemma about an invariant extension of a subset in a multimetric space for a Lipschitz map is stated. In Section **5** a solution of the value prediction problem is discussed.

2. A quick review about categories and multimetric spaces

A *category* is a class $\mathcal{C}$ of objects together with a family of sets $Hom(A, B)$, one for each pair of objects A, B, equipped with a notion of composition $Hom(A, B) \times Hom(B, C) \to Hom(A, C) : (f, g) \mapsto g \circ f$, enjoying the same properties as the set-theoretical maps in the category **Set** of sets. The elements of $Hom(A, B)$ are called *morphisms* or *arrows* from A to B. In the category **Set** of sets the objects are sets and the morphisms are maps of sets. The morphisms of a category $\mathcal{C}$ obey two laws. The former is the *identity law*, i.e there exists an identity arrow $1_A \in Hom(A, A)$ such that $f \circ 1_A = f$, $1_A \circ g = g$ for any two arrows f, g for which the composition is defined. The latter is the *associativity law* $f \circ (g \circ h) = (f \circ g) \circ h$ for any three composable arrows f, g, h. For a nice introductory and elementary account on the use of categories in Geometry see Gatto (2000).

Mathematical theories study objects and morphisms of one or more interrelated categories. For example, functional analysis considers categories of linear spaces, metric spaces, measure spaces, etc., depending on the problem.

Categories themselves form a category **Cat**, with categories as objects and functors as arrows. A *functor* $F : \mathcal{C} \to \mathcal{D}$ is a map of the objects of $\mathcal{C}$ to the objects of $\mathcal{D}$ and of the morphisms of $\mathcal{C}$ to the morphisms of $\mathcal{D}$, such that $F(1_A) = 1_{F(A)}$, where A is an object in $\mathcal{C}$, and $F(f \circ g) = F(f) \circ F(g)$, where f, g are composable arrows in $\mathcal{C}$. For example, the law assigning to each vector space the dual space of linear forms, $(-)^* : \mathbf{Vect} \to \mathbf{Vect}$, is a contravariant functor, which associate to

each linear map its transpose. The functor is contravariant because $(g\circ f)^* = f^*\circ g^*$, *i*.e., it changes the direction of the arrows, and the composition law to the opposite ones.

A multimetric space is a set M equipped with a set of metrics $\mathcal{O}_M$. Multimetric spaces form a category **MultiMet** whose objects are multimetric spaces, and whose arrows are maps $f : (M, \mathcal{O}_M) \to (N, \mathcal{O}_N)$, where $f : M \to N$ is such that the induced map of metrics

$$\begin{cases} f^* & : & \mathcal{O}_N & \longrightarrow & \mathcal{O}_M \\ & & d_N & \longmapsto & f^*d_N := d_N \circ f \times f \end{cases}$$

is well defined, that is $f^*(\mathcal{O}_N)$ is indeed a set of $\mathcal{O}_M$.

Metrics allow to *capture* a class of relevant transformations in the sense that these transformations preserve the distinguished sets of metrics. In the following, a *data set* will be an additional geometric object attached to the set of metrics. They bring together a joint invariant, the extension of data.

3. Problem statement and categorical model of data

Data will be geometrically modeled by subsets of points of a multimetric space (See Section **2**). In this case, the metrics are chosen in such a way that are preserved by the selected class of transformations which, depending on the needs of the problem, do not alter the 'information contained in the data'. For example:

(1) a set of points in the plane with a given density can be considered equivalent to the original one if it is rotated or translated;
(2) a data table with some repeated columns can be thought of as providing the same information of the original smaller one, but presented in a higher dimensional space.

The metric for the first case is standard Euclidean, for the second case is the set of Euclidean metrics, induced by embedding from the standard Euclidean one.

It is convenient to assume that there is a partition/foliation on the multimetric space expressing constraints on the data points. For example, the graph of a function in a Euclidean space can be considered foliated by the graphs of the functions induced by the original one by keeping constant some variables.

The problem is to extend the original set of data into leaves of the foliation disjoint from the original set such that the operation of extension was natural, commuting with the accepted transformations of data.

Definition 1. *A data model is a category* **MD**, *consisting of quadruples* $(M, \mathcal{O}_M, S, D)$, *where* M *is a set,* $\mathcal{O}_M$ *is a set of metrics on* M, S *is a foliation on* M, D *are data in* M. *A morphism in* **MD** *is a map* $f : M \to M'$ *such that for any metric* $\rho_{M'} \in \mathcal{O}_{M'}$ $f^*(\rho_{M'}) := \rho_{M'} \circ f \in \mathcal{O}_M$, *for any leaf* $s \in S$, $f(s) \subset s'$, $f(D) \subset D'$.

Among all the morphisms in **MD** there is a subcategory $\mathbf{MD_{inf}}$, said to be of infomorphisms, with the same objects as in **MD** but fewer arrows, which, informally speaking, preserve information/identification of the data in the sense as was indicated

above. $\mathbf{MD_{inf}}$ is a subcategory of $\mathbf{MD}$ obtained by taking a selected subset of arrows. Infomorphisms constitute somewhat like a priory information in the statistical approach, but taken from a different consideration. Depending on the problem, the category of infomorphisms $\mathbf{MD_{inf}}$ can vary. In most cases it is generated by Euclidean motions, embeddings to higher dimensional Euclidean spaces by duplicating some parameters, and the other identification morphisms, expressing the similarity of the data representations in the problem. The category $\mathbf{MD}$ has a natural projection p onto the category $\mathbf{M}$, in which the last element of the quadruple (the data D) is forgotten. $p : \mathbf{MD} \to \mathbf{M}$ is a fibred category over the base $\mathbf{M}$ with the Cartesian morphisms determined by pullbacks in $\mathbf{M}$ (about fibred categories cf. Jacobs (1999) and Kondratiev (2006)). Similarly, $\mathbf{M}$ has a projection onto the category of multimetric spaces $\mathbf{MultiMet}$, which forgets the partition/foliation on the space.

Definition 2. *An extension of the data is an endofunctor*

$$F : \mathbf{MD_{inf}} \to \mathbf{MD_{inf}}$$

of the subcategory $p \circ i : \mathbf{MD_{inf}} \to \mathbf{M}$ *of the fibered category* $p : \mathbf{MD} \to \mathbf{M}$*, where* $i : \mathbf{MD_{inf}} \hookrightarrow \mathbf{MD}$ *is the inclusion functor, together with a vertical natural transformation* $\varepsilon : i \Rightarrow i \circ F$ *with the components not being reduced to the infomorphisms, that is*

$$p \circ i \circ F = p \circ i,$$
$$p\varepsilon = 1_{poi},$$

and $\varepsilon_{(M,\mathcal{O}_M,S,D)}$ *is an arbitrary morphism of the category* $\mathbf{MD}$*.*

Proposition 1. *Extensions of the data form a monoid with respect to the operation* $\eta \odot \varepsilon = (\eta * 1_F) \circ \varepsilon : i \Rightarrow i \circ G \circ F$*, where* $\eta : i \Rightarrow i \circ G$*,* $\varepsilon : i \Rightarrow i \circ F$ *are any two extensions,* $*$ *and* $\circ$ *are respectively the horizontal and vertical compositions of natural transformations.* $1_i : i \Rightarrow i$ *is the neutral element of the monoid.*

If, at least, one extension ε is known then ε^n is also an extension for $n \in \mathbb{N}$, possible coinciding with the original one.

For statistical data it makes sense to take into account such extensions ε, which preserve density of the data, that is, if $D_1 \approx D_2$ as having approximately the same density, then $\varepsilon(D_1) \approx \varepsilon(D_2)$.

In Section **4** a construction of a nontrivial extension is given in the case when the admissible transformations are taken to be isomorphisms.

4. Lemma about an invariant extension

In this section Lipschitz maps of multimetric spaces are considered as a tool to express uncertainty of the disposition of the points of a set in the space with respect to each other.

A map of metric spaces $f : (M, d_M) \to (N, d_N)$ is Lipschitzian with constant L, if for any pair of points $(x, y) \in M \times M$ the inequality $d_N(f(x), f(y)) \le L \cdot d_M(x, y)$ holds true. In the case, when f satisfies the condition $d_N \circ f \times f = d_M$, it is automatically Lipschitzian with the constant 1. The morphisms in the category of

data **MD** are by definition Lipschitz maps with the constant 1 for any pair of the corresponding metrics $\mathcal{O}_N \ni d_N \mapsto d_N \circ f \times f \in \mathcal{O}_M$.

Definition 3. *A Lipschitz uncertainty $e(f, m, D)$ of the map of multimetric spaces $f : (M, \mathcal{O}_M) \to (N, \mathcal{O}_N)$ at the point $m \in M$ with respect to the set $D \subset M$ is the diameter of the intersection $\bigcap_{d_N \in \mathcal{O}_N} \bigcap_{d \in D} B_{f(d), d_N(f(m), f(d))}$, where $B_{f(d), d_N(f(m), f(d))}$ is a ball in the space N for the metric d_N with the center $f(d)$ and radius $d_N(f(m), f(d))$, $m, d \in M$. The diameter of the set X is, by definition, a nonnegative number (including $+\infty$)*

$$\delta(X) := \inf_{d_N \in \mathcal{O}_N} \sup_{(x_1, x_2) \in X^2} d_N(x_1, x_2).$$

As it follows from the definition, $e(f, m, D) \geq e(1_M, m, D)$ for any map of multimetric spaces f, since

$$\begin{aligned} e(f, m, D) &= \delta\Big(\bigcap_{d_N \in \mathcal{O}_N} \bigcap_{d \in D} B_{f(d), d_N(f(m), f(d))} \Big) \geq \\ &\geq \delta\Big(\bigcap_{d_M \in f^*(\mathcal{O}_N)} \bigcap_{d \in D} B_{d, d_M(m, d)} \Big) \geq \\ &\geq \delta\Big(\bigcap_{d_M \in \mathcal{O}_M} \bigcap_{d \in D} B_{d, d_M(m, d)} \Big) = e(1_M, m, D), \end{aligned}$$

where $f^*(\mathcal{O}_N)$ is the image of O_N via the induced map of metrics

$$f^* : d_N \mapsto d_N \circ f \times f.$$

If f is an isomorphism, then $e(f, m, D) = e(1_M, m, D)$.

Lemma 1.

(a) *If $f : M \simeq N$ is an isomorphism in the category* **MD** *and $g : N \to K$ is a morphism in the category* **MultiMet**, *then $e(g \circ f, m, D) = e(g, f(m), f(D))$.*

(b) *If, conversely, $f : M \to N$ is an arbitrary morphism in the category* **MD** *and $g : N \simeq K$ is an isomorphism in the category* **MultiMet**, *then $e(g \circ f, m, D) = e(f, m, D)$.*

Proof. As for item (a) one has the equality

$$e(g \circ f, m, D) = \delta\Big(\bigcap_{d_K \in \mathcal{O}_K} \bigcap_{d \in D} B_{g \circ f(d), d_K(g \circ f(m), g \circ f(d))} \Big) =$$

$$= \delta\Big(\bigcap_{d_K \in \mathcal{O}_K} \bigcap_{f(d) \in f(D)} B_{g(f(d)), d_K(g(f(m)), g(f(d)))} \Big) = e(g, f(m), f(D)).$$

For item (b)

$$e(g \circ f, m, D) = \delta\Big(\bigcap_{d_K \in \mathcal{O}_K} \bigcap_{d \in D} B_{g \circ f(d), d_K(g \circ f(m), g \circ f(d))} \Big) =$$

$$= \delta\Big(\bigcap_{g^*(d_K) \in \mathcal{O}_N, d_K \in \mathcal{O}_K} \bigcap_{d \in D} B_{f(d), g^*(d_K)(f(m), f(d))} \Big) =$$

$$= \delta(\bigcap_{d_N \in \mathcal{O}_N} \bigcap_{d \in D} B_{f(d),d_N(f(m),f(d))}) = e(f,m,D)$$

(in the second equality the fact that the diameter of a set δ is invariant under isomorphisms of multimetric spaces is used). □

Lemma 2 (invariant extension)**.** *Let $\mathcal{M} = (M, \mathcal{O}_M, S, D)$ be an object of the category* **MD** *and $h : (M, \mathcal{O}_M) \to (K, \mathcal{O}_K)$ be a map of multimetric spaces. For each leaf $s \in S$ there exists a subset $p_s \subset s$ invariantly attached to the pair $(\mathcal{M}, h) \in Ob(\mathbf{MD} \times \mathbf{MultiMet}^{\to})$ with respect to the isomorphisms $(f, g) \in Ar(\mathbf{MultiMet}^{\to})$, where f is also an arrow in* **MD**, *that is for any commutative diagram*

$$\begin{array}{ccc} (K, \mathcal{O}_K) & \xrightarrow[\sim]{g} & (K', \mathcal{O}_{K'}) \\ {\scriptstyle h}\uparrow & & \uparrow{\scriptstyle h'} \\ (M, \mathcal{O}_M, S, D) & \xrightarrow[f]{\sim} & (M', \mathcal{O}_{M'}, S', D') \end{array}$$

and each leaf $s \in S$, the equality $f(p_s) = p_{f(s)}$ holds. The set p_s is to be defined as a set of minimum points of the function $e(h, m, D)$ with respect to the point $m \in M$ under the constraint $m \in s$ for each $s \in S$.

Remark. Notation $\mathbf{MultiMet}^{\to}$ stands, as usual, for the *category of arrows* over the base category **MultiMet**. Its objects are the arrows of **MultiMet** and its morphisms are the commutative squares of arrows in **MultiMet** similar to the above diagram.

Proof. From Lemma 1 it follows that

$$e(h, m, D) = e(g \circ h, m, D) = e(h' \circ f, m, D) = e(h', f(m), f(D))$$

for any point $m \in M$. In particular, for any leaf $s \in S$, the level sets of the functions $e(h, m, D)$ and $e(h', m', D')$, where $m \in s$, $m' \in f(s)$, are in one-to-one correspondence via the map f. Consequently, the sets of minimal points $p_s \subset s, p_{f(s)} \subset f(s)$ of these functions are in one-to-one correspondence via the map f, that is for any leaf $s \in S$ the equality $f(p_s) = p_{f(s)}$ holds as well. □

Remarks.

- The size of the set of minimal points p_s depends on the choice of the set of metrics in the space. Obviously, p_s is always nonempty. It can be a one point set, discrete, continuous or any other one. For 'appropriate' metrics it is usually a one-point or discrete set independently on the leaf $s \in S$.
- If $s \cap D \neq \emptyset$, then $p_s \supset s \cap D$, that is $\coprod_{s \in S} p_s$ is an extension of the data D.
- Although $e(h, m, D) \geq e(1_M, m, D)$ for any map h of multimetric spaces and, consequently, $\min\{e(1_M, m, D) \,|\, m \in s\} \leq \min\{e(h, m, D) \,|\, m \in s\}$, it does not mean that the points of minimum are the same for both functions. Nevertheless, the choice of the identity arrow $1_{\mathcal{M}}$ for each object $\mathcal{M}$ of the category **MD**, commuting with any (iso)morphism f, reduces to zero the number of arbitrary parameters of the construction, making it canonical.

5. Prediction of empirical data

A direct use of the results of Section **4**, that is the prediction of a new case out of a population based on the criterium of minimum Lipschitz uncertainty for the identity map $e(1_M, m, D) \xrightarrow[m\in s, s\in S]{} \min$, works and shows results better than those obtained from the regression or statistical technique. In particular, the method predicts with a 'uniformly' small individual error for each case almost independently on if it is close to the average or extremal.

The algorithm of prediction was tested for the price evaluation in the real estate market of the Montreal city, Canada. The individual errors were: for condominiums within $1-2\%$, and for single family houses within $5-10\%$, which is much better than in the widely used in the real estate Vandell's method (Vandell 1991; Gau, Lai, and Wang 1992; Lai *et al.* 2008), based on statistical estimation. Note, that in the real estate area a method of prediction is considered accurate if the *average* error is within 10%. In extreme cases the individual error is allowed to be the hundreds of percent. In this sense the known statistical methods are far from being uniformly predictive in opposition to the proposed one.

This phenomenon can be explained as follows. Accuracy of the method is a consequence of little and subtle model assumptions, so that the prediction does not contain systematic model errors. Uniformity is provided with the main condition, that the inclusion "sample $\subset$ population" is internal in the sense of Section **3**. This resembles the point of view of the older statisticians, that the membership relation "case $\in$ population" is determined by the same condition independently on the case, that is uniformely. The matter of this fact is essentially categorical, so before the category theory had been created it was hardly feasible to work with it in an operational way.

It is possible to modify the method to work directly with the density, the main 'visual' characteristic of the data. In this approach, the point of the prediction will be a point at which Lipschitz uncertainty is minimal, that is one at which the density coordinate is the most reliably calculated within the framework of the method.

The modification is as follows. In the object $(M, \mathcal{O}_M, S, D)$ the partition or foliation S is being forgotten and the result is being extended to the object $(\mathbb{R}^+ \times M, \mathcal{O}_{\mathbb{R}^+\times M}, \hat{S}, \hat{D})$, where $\hat{S}$ consists of the fibres of the projection

$$\mathbb{R}^+ \times M \to M,$$

the new data set is

$$\hat{D} = \{(p_m, m) \,|\, m \in D,\ p_m \text{ is the density of data } D \text{ at point } m\},$$

the class of metrics is

$$\mathcal{O}_{\mathbb{R}^+\times M} = \{|p_{m_1} - p_{m_2}| + d_M(m_1, m_2), (|p_{m_1} - p_{m_2}|^2 + d(m_1, m_2)^2)^{\frac{1}{2}} \,|$$

$$d_M \in \mathcal{O}_M, (p_{m_1}, m_1), (p_{m_2}, m_2) \in \mathbb{R}^+ \times M\}.$$

The class of metrics is selected in order to include only shifts on $\mathbb{R}^+$ as acceptable transformations of the density coordinate.

The density at the data points of a multimetric space can be found in the following way. Assume that the cardinality of the data set D is equal to N. For each metric $d_M \in \mathcal{O}_M$ choose two distances r_{d_M} and R_{d_M} between the closest and the most far points from D. Let $\varepsilon > 0$ be a small number. Then for each function of the type $f : \mathcal{O}_M \to \mathbb{R}^+$ pointwise placed between functions $\alpha : d_M \mapsto r_{d_M} - \varepsilon$ and $\omega : d_M \mapsto R_{d_M} + \varepsilon$ and for each data point $m \in D$ there is a number N_m of the data points inside of the set $\bigcap\limits_{d_M \in \mathcal{O}_M} B_{m,f(d_M)}$. So that, there exists a map

$$(f, m) \mapsto p_m = \frac{N_m}{\sum\limits_{m \in D} N_m},$$

specifying for each f the density $p(f) : D \to \Delta^{N-1}$ at the data points D, where Δ^{N-1} is the standard simplex of dimension $N-1$. Then the cardinality of the image of $p(f)$ is $\leq N$ and the cardinality of the union $\bigcup\limits_{\alpha<f<\omega} \mathtt{Im}(p(f))$ is smaller than N^N. The choice of a point from $\bigcup\limits_{\alpha<f<\omega} \mathtt{Im}(p(f))$ with the highest number of different coordinates gives the most appropriate estimation of the density of data D. The number of the functions f can be made finite if, for example, only 'parallel' functions of the form $f_k : d_M \mapsto \alpha(d_M) + k \cdot h_{d_M}$, $k = 0, \ldots, K$, $K \cdot h_{d_M} = \omega(d_M) - \alpha(d_M)$, being shifted from one another for each argument d_M by the step h_{d_M} are taken into account. This, however, can bring just a quasioptimal estimate of the density.

5.1. A practical algorithm of prediction. For many cases the following procedure will be enough. In the parameter space $M \simeq \mathbb{R}^n$ (or $M \simeq I^n$, where $I = [0, 1]$), containing the data D,

Step 1. *choose a set of metrics $\mathcal{O}_M$, such that the group of automorphisms $\mathbf{Aut}(M, \mathcal{O}_M)$ of the multimetric space $(M, \mathcal{O}_M)$ would transform the data D to other data, which are supposed to contain the same information.*

Comment to Step 1. For many cases the set of metrics $\mathcal{O}_M$ will contain either one standard Euclidean metric d_{std}, ensuring that the standard Euclidean density of the data will be preserved, or a couple of metrics

$$|y_1 - y_2| + d_{std\setminus y} \qquad \text{and} \qquad (|y_1 - y_2|^2 + d^2_{std\setminus y})^{1/2},$$

where y is the unique dependent variable, $d_{std\setminus y}$ is the standard Euclidean metric on the subspace of independent variables, ensuring that the standard Euclidean density of the data will be preserved and the transformations of dependent variable y will be just shifts. However, sometimes the choice of the set $\mathcal{O}_M$ can be more complicated.

Step 2. *Form the Lipschitz uncertainty function $e(1_M, m, D)$ for the point of prediction $m = (y, x) \in M$, where y is the dependent variable, x are independent variables (x are known, y is unknown).*

Comment to Step 2. By definition, $e(1_M, m, D) := \delta(\bigcap\limits_{d_M \in \mathcal{O}_M} \bigcap\limits_{d \in D} B_{d,d_M(m,d)})$. In this case the function $e(1_M, m, D)$ depends on one real variable y.

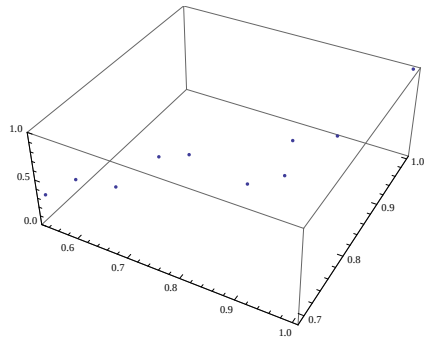

FIGURE 1. The original data.

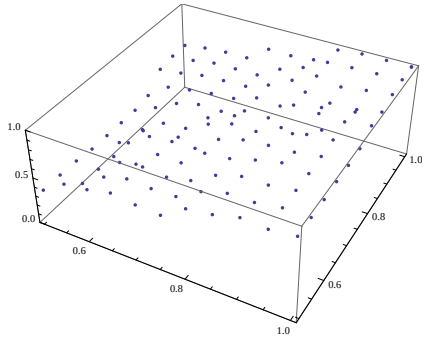

FIGURE 2. Extension of the original data with 100 new points.

Step 3. *Find the global minimum point/points* (y_{min}, x) *of the function* $e(1_M, m, D)$. *This is the one-point prediction.*

Example (Cf. Vandell (1991))**.** *The original data are 10 family houses determined by 4 parameters (living area in square feet, lot size in square feet, number of bathrooms and number of bedrooms).*

Price ($)	*Living Area*	*Lot Size*	*Bathrooms*	*Bedrooms*
180000	2320	12000	2	3
145400	2200	11200	1.5	3
240000	2950	11600	2	4
263000	3020	12800	2.5	4
289500	3370	15600	4.5	5
93400	1800	10800	1	2
249000	3200	13400	3	4
136700	1950	11100	2	2
249300	3100	12000	3.5	4
205500	2500	12100	2.5	3

For a graphical 3D-illustration of how the method works, we restrict ourselves to the first two independent parameters (living area and lot size) which are the most important for the prediction of the price.

Figure 1 represents the (normed) original data. Figure 2 shows the result of a discrete extension of the original data with 100 new points. The surface in figure 3 is a smooth interpolation of the extension. The proposed algorithm was implemented in Wolfram Mathematica 8.0, by using the standard Euclidean metric. Figures 4 and 5 illustrate the behavior of Lipschitz uncertainty as a function of price at the points $(1, 0)$ *and* $(0.7, 0.7)$ *of the parameter space. The values of the price prediction are the global minimum points of the functions. The practical implementation of the method is a global minimization of a function of one variable at each point of the parameter space. The minimized function* $e(1_M, m, D)$ *is constructed following definition 3.*

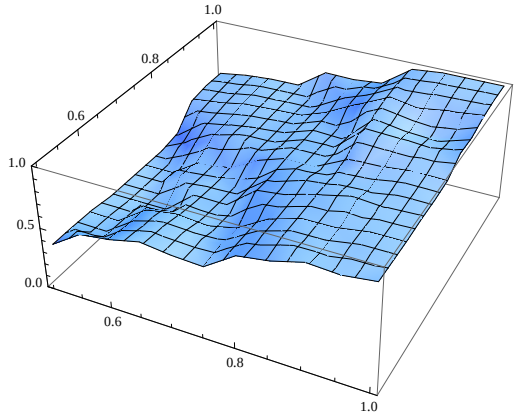

Figure 3. Interpolation of the extended data.

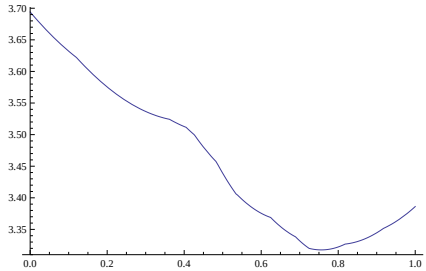

Figure 4. Lipschitz uncertainty $e(1_M, m, D)$ as a function of price at the point $(1, 0)$.

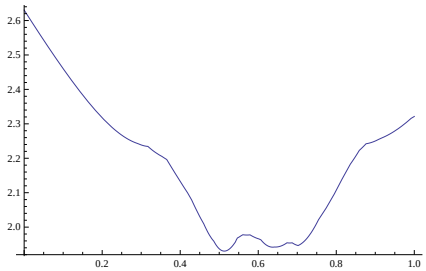

Figure 5. Lipschitz uncertainty $e(1_M, m, D)$ as a function of price at the point $(0.7, 0.7)$.

6. Conclusion

The paper is an attempt to introduce the idea of invariance, which is widely exploited in physics and geometry, to the area of prediction. Another goal was to show that the Lipschitz condition fits well into the sphere of computational methods and is the key condition for the invariant extension construction. It may be that there are other invariant extensions of the data that do not meet the criterion of minimum Lipschitz uncertainty of the identity map, but they are not yet known.

The method was tested in predicting real estate prices, where it showed better and more stable results compared to the known ones (Vandell 1991; Gau, Lai, and Wang 1992; Lai *et al.* 2008). Some explanation may be given with a slogan 'the more a thing is invariant, the more chances it has to exist objectively'. Indeed, the extension of a sample to the real population is definitely invariant with respect to the acceptable representations of them in the context of the problem. The described method was created to not only approximate the empirical data, but to try to

capture the real manifestation of the process, based on the invariant properties of the data.

Acknowledgments

I am deeply grateful to L. Gatto for a fruitful discussion on the subject presented in the paper.

References

Gatto, L. (2000). *Intersection Theory on Moduli Space of Curves*. Vol. 61. Monografias Matemática do IMPA. Rio de Janeiro: Instituto Nacional de Matemática Pura e Aplicada. URL: http://www.impa.br/opencms/pt/biblioteca/mono/Mon_61.pdf.

Gau, G. V., Lai, T., and Wang, K. (1992). "Optimal Comparable Selection and Weighting in Real Property Valuation: An Extension". *Journal of the American Real Estate and Urban Economics Association* **20** (1), 107–123. URL: http://www.areuea.org/publications/ree/view_article.phtml?id=6489.

Jacobs, B. (1999). *Categorical Logic and Type Theory*. Ed. by S. Abramsky, S. Artemov, R. A. Shore, and A. S. Troelstra. Vol. 141. Sudies in Logic and the Foundations of Mathematics. Amsterdam: Elsevier.

Kondratiev, G. V. (2006). "Manifolds, Structures Categorically". arXiv: math/0608503.

Lai, T., Vandell, K., Wang, K., and Welke, G. (2008). "Estimating Property Values by Replication: An Alternative to the Traditional Grid and Regression Methods". *Journal of Real Estate Research* **30** (4), 441–460. URL: http://aux.zicklin.baruch.cuny.edu/jrer/papers/abstract/past/av30n04/vol30n04_03.htm.

Vandell, K. D. (1991). "Optimal Comparable Selection and Weighting in Real Property Valuation". *Journal of the American Real Estate and Urban Economics Association* **19** (2), 213–239. URL: http://www.areuea.org/publications/ree/view_article.phtml?id=5124.

* Nizhny Novgorod State Technical University, named after R. E. Alekseev, Russia

Email: gennadii.kondratiev@gmail.com

Communicated 28 November 2013; published online 14 December 2013