Topological Data Analysis to improve the predictive model of an Electric Arc Furnace

MATTIA DE COLLE
Abstract

Data mining, and in particular topological data analysis (TDA), had proven to be successful in abstracting insights from big arrays of data. This thesis utilizes the TDA software Ayasdi™ in order to improve the accuracy of the energy model of an Electric Arc Furnace (EAF), pinpointing the causes of a wrong calculation of the steel temperature. Almost 50% of the charges analyzed presented an underestimation of temperature, while under 30% an overestimation.

First a dataset was created by filtering the data obtained by the company. After an initial screening, around 700 charges built the dataset, each one characterized by 104 parameters. The dataset was subsequently used to create a topological network through the TDA software. By comparing the distribution of each parameter with the distribution of the wrong temperature estimation, it was possible to identify which parameters provided a biased trend. In particular, it was found that an overestimation of temperature was caused by an underestimation of the melting energy of materials not having through a melting test. It was also found a possible biased trend in some distribution of parameters like %O in steel and slag weight, which it is believed are all connected together. Despite not finding a global solution for the reasons behind the underestimation of temperature, it is believed that a different settings more focused around the materials used as scrap mix can highlight more on that subject. In conclusion TDA proved itself efficient as a problem solving technique in the steel industry.
Table of Contents

1. Introduction ....................................................................................................................... 1
2. Background ....................................................................................................................... 3
   2.1. The Electric Arc Furnace process .............................................................................. 3
   2.2. Slag and slag basicity ............................................................................................... 5
   2.3. Temperature measurements in Electric Arc Furnace ............................................... 6
   2.4. Energy models in Electric Arc furnace ...................................................................... 7
   2.5. Topology and Topological Data Analysis ................................................................. 8
   2.6. TDA using Ayasdi™ ................................................................................................. 9
3. Method ............................................................................................................................ 13
   3.1. Problem formulation ............................................................................................... 13
   3.2. Estimation of gains .................................................................................................. 15
   3.3. Database and network construction ........................................................................ 16
   3.4. Hypothesis formulation ......................................................................................... 17
4. Results ............................................................................................................................ 19
   4.1. The network .......................................................................................................... 19
   4.2. Confronting the color coding ................................................................................. 20
5. Discussion ....................................................................................................................... 30
   5.1. Flare study ............................................................................................................. 32
6. Conclusions ..................................................................................................................... 34
7. Further work .................................................................................................................. 36
Bibliography ....................................................................................................................... 37
**List of Abbreviations**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EAF</td>
<td>Electric Arc Furnace</td>
</tr>
<tr>
<td>TDA</td>
<td>Topological Data Analysis</td>
</tr>
<tr>
<td>$T_c$</td>
<td>Calculated Temperature ($^\circ C$)</td>
</tr>
<tr>
<td>$T_m$</td>
<td>Measured Temperature ($^\circ C$)</td>
</tr>
<tr>
<td>$\Delta T$</td>
<td>$T_c - T_m$, Delta Temperature ($^\circ C$)</td>
</tr>
<tr>
<td>OPC</td>
<td>Over Predicted Charge</td>
</tr>
<tr>
<td>UPC</td>
<td>Under Predicted Charge</td>
</tr>
<tr>
<td>APC</td>
<td>Accurately Predicted Charge</td>
</tr>
<tr>
<td>B</td>
<td>Basicity</td>
</tr>
<tr>
<td>$Q$</td>
<td>Heat absorbed by scrap ($J$)</td>
</tr>
<tr>
<td>$c_p$</td>
<td>Specific heat capacity at constant pressure ($\frac{J}{kg^\circ C}$)</td>
</tr>
<tr>
<td>$m$</td>
<td>Mass of scrap ($kg$)</td>
</tr>
<tr>
<td>$T_{\text{melt}}$</td>
<td>Melting temperature of scrap ($^\circ C$)</td>
</tr>
<tr>
<td>$T_{\text{preheat}}$</td>
<td>Scrap temperature after the preheating phase ($^\circ C$)</td>
</tr>
<tr>
<td>$L_f$</td>
<td>Latent heat of fusion ($\frac{kJ}{kg}$)</td>
</tr>
<tr>
<td>$E_{d,\text{calc}}$</td>
<td>Calculated energy demand ($J$)</td>
</tr>
<tr>
<td>$E_{d,\text{real}}$</td>
<td>Real energy demand ($J$)</td>
</tr>
</tbody>
</table>
1. Introduction

The Electric Arc Furnace (EAF) is a growing metallurgical process, through which metal scrap is converted into new steel. Scrap is melted through electric arcs which generates an unstable and intense environment inside the furnace. This translates into high productivity but a low predictability. Though one of the key aspects is being able to predict the temperature inside the furnace, which usually dictates the process time. Frequent temperature measurements inside the furnace, which often reaches 1500-1600°C, are hard to accomplish and not economically viable. Therefore, being able to calculate the temperature through a predictive model is often desirable. Today, the temperature is commonly calculated using an energy balance, including fundamental energy requirements of the inserted materials, together with the electrical energy and estimations of system losses. However, due to the unstable environment, the predictability of such models is very low. Hence, there is a strong drive to increase the accuracy of these models to achieve lower energy consumption and process time.

It is not unusual for companies nowadays to amass vast amount of complex data of their processes [1]. This provides tremendous opportunity for improvement and discovery of new insights on the subjects involved, as well as having the chance to generate an incredible value in terms of both optimization and revolution of the procedures in use. Although dealing with big arrays of data is not an easy task especially with traditional ways of problem solving [2]. A conventional analytic approach is hypothesis-driven, through which multiple claims are tested using the data available. This method is both time consuming and expensive and does not provide the certainty of a result. Reiteration of multiples queries does not assure any kind of success, especially because the chances decrease the more complex the data grows. The necessity for a better and more systematic approach to big dataset caused the rise of a new branch of analysis, called Data Mining. This method groups different techniques of pattern extrapolation, like machine learning, artificial intelligence and statistics. One of the many options available with Data Mining is Topological Data Analysis (TDA), which studies the geometry of a data set, and through various algorithms produces insights often invisible to other conventional methods.
This project focused on improving the accuracy of the EAF predictive model of Outokumpu’s steel plant in Avesta (Sweden). Analyzing the differences between meltdowns where the temperature is accurately calculated with the ones where it is not, will highlight which parameters of the predictive model must be revised to increase its accuracy. This project aims to use the TDA software Ayasdi™, comparing measured and calculated temperature, along with data from the inserted materials and the furnace system, in order to find patterns that will lead to a classification of multiple meltdowns and their temperature predictability.
2. Background

2.1. The Electric Arc Furnace process

The importance of the EAF relies on the necessity for sustainability in the steel manufactory industry. In fact, it is estimated that some decades from now, due to the increase of steel demand by emerging countries, the resource scarcity will make it impossible to keep up with the growing demand, causing a consequent raise in price [4]. Currently, the EAF is the only steelmaking process that melts metal scrap to obtain new steel, relying on recycled steel instead of using iron ore. Roughly 70% of the world steel production is made by conventional processes [5], but it is estimated that through the years the usage of the EAF will increase [6].

Unlike the basic oxygen route [7], which produces steel starting from raw materials, the EAF charges recycled material, that are heated up by circulating high currents inside the materials, causing an increase in temperature above their melting point. In most cases, and also the one under study by this project, the current is powered by a three-phase electrical supply. The current is then transmitted to the scrap by three graphite electrodes that are lowered down into the furnace from the roof top (as is possible to see in Figure 2).

![Figure 2 Cross section of an Electric Arc Furnace](image-url)
Since the EAF is a batch process there are several repetitive steps which define a cycle often called “tap-to-tap”, which is commonly around sixty minutes long. The schedule is usually set as follows:

- Charging the first basket
- Initial melting period
- Charging the second basket
- Second melting period
- Refining
- Tapping
- Maintenance before next cycle

It is worth to notice that the number of baskets used may vary significantly according to the steel plants necessities. The previous process description used an arbitrary number two just to show that all the baskets are charged before the refining takes places. Also scrap, before being lowered down, inside the furnace is preheated with exhaust gases from previous cycles in order to remove the moist (which can cause explosions) and decrease the energy needed to reach the melting point.

The ultimate EAF goal is to provide the maximum amount of energy in the shortest time possible, while keeping to minimum energy losses, refractory wear and electrode consumption. Due to wear of the refractory walls, at the beginning of the melting process the voltage is kept low until the electrodes bore into the scrap pile. This is done in order to prevent excessive radiation caused by an uncovered arc, which increase the damage to the furnace walls substantially. Once the arc is covered by unmolten scrap, the voltage is increased and the furnace is operating at maximum power. When there is an appropriate percentage of molten scrap the next basket is charged inside and subsequently the lancing operations can start. From this brief explanation it is clear how complex in terms of analysis this system is. There are hundreds of variables that need to be taken into account to improve the model’s accuracy.
2.2. Slag and slag basicity

Slag is a fundamental part of the steel making process, and it is in this project’s interest to introduce some of the key aspects about it in order to fully comprehend a later part of this report. Slag is by definition “an ionic solution consisting of molten metal oxides that float on top of the steel (completely liquid or partially liquid)” [9]. In fact, slag forms naturally during the melting phase of the scrap, and due to a lower density than molten steel, floats on top of the steel bath. Despite the formation of slag comes naturally, there is a great interest in manipulating the process in order to obtain the right composition. In fact, a proper slag composition is key to achieve a high quality steel. Slag covers many roles in the steel making process [9]:

- Covers the bottom of the electrodes protecting the refractories from the arc radiations.
- Improves the quality of steel absorbing unwanted elements and oxides.
- Protects the steel from oxidation.
- Substantially decreases the absorption of nitrogen and hydrogen.
- Insulate the steel bath reducing the heat losses.

Slag viscosity is also a crucial parameter. High fluidity and low viscosity is desired to accelerate the exchange of materials between slag and the metal bath while being detrimental to the refractories increasing the wear.

Slag basicity: similar to the concepts of pH slag basicity comes in handy as a universal parameter that could define each kind of slag despite the high variance in composition [9]. In fact, slag is expressed by its many oxides weight percentage and without a global parameter it would be hard to compare two different slags. Oxides are then classified as acid, basic or amphoteric (they may change their behavior depending on the surrounding conditions) [9].

The most common approach to slag basicity is the B ratio defined as follows [9]

\[ B = \frac{\%CaO}{\%SiO_2} \]

Eq1.

Slag basicity assure that slag has the proper viscosity and refining capability.
As a first approximation Eq1 is a good solution but several problems lie in adopting such a definition. Although it is not in this project’s goal to dig into the matter, but rather give a basic background to comprehend further work.

2.3. **Temperature measurements in Electric Arc Furnace**

The strong drive in improving the EAF’s process pushed the development of new and innovative techniques granting huge gains in productivity. Although despite being one of the most critical factors, the state of the art of temperature measurements remains untouched [10]. In fact, there is a large variation between the progress of individual melting cycles due to the different nature of the scrap used and the thermal efficiency of the furnace [11]. The scrap’s heat capacity varies substantially based on the kind of materials inserted, which makes melting times and temperatures hard to control. Therefore, an accurate temperature estimation system is mandatory to define the correct end point of the melting stage [11].

Nowadays the common equipment for temperature measurements is an automated lancing system, where a disposable thermocouple placed on a probe holder is dipped into the steel bath [10, 11]. This process proved itself to not be reliable enough to guarantee problem-free tapping for every charge. It is common to observe erratic temperature profiles alongside the presence of solid scrap near to end of the melting cycle [11]. There are several concerns regarding the lancing system commonly used nowadays: first of all, the measurement frequency is quite low (around 60 seconds between dips), there’s a high risk of failure due to the fact that the probe is dipped into an inhomogeneous environment, and the current methods represent also a dangerous place for the operator [11]. A possible way to increase efficiency is using automatic lancing robot. In this way operability is increased substantially both in safety for the operator and in terms on process optimization.

Several solutions had been considered in order to replace the temperature measurements by probing. Continuous thermocouples either in the furnace walls or attached to the furnace roof, as well as infrared sensors pointed at the steel bath have been investigated as possible substitutes. A different route consists in building predictive models (as this project’s study case) to derive the temperature from a mass and heat balance, or by exploiting other parameters like vibrations.
or harmonic distortion on the power grid [10]. Although none of these systems proved to be more reliable than the common state of art, so more work on reliability of these new systems needs to be done.

2.4. **Energy models in Electric Arc Furnace**

The low predictability of the EAF, especially regarding the random behavior of the electric arcs, the quality and quantity of steel inserted and size of the furnaces make it hard to build a predictive model able to simulate the process’ progress [12]. Due to this reason, computer models adaptable to a huge variety of EAFs are hard to develop, forcing companies to tailor one to their specific steel plant’s case. Although, it seems that it is common practice to adopt a model based on a mass and heat balance, even though the complexity may vary as well as the fundamental principles used as hypothesis [13]. For instance, it is possible to fraction the EAF furnace volume between zones that exchange mass and heat between themselves or each other [13]. This division helps modeling the behavior of the melting scrap and the various reactions along the all process.

The same principles described above are applied for this project’s control system. The software with several parameters calculates a mass and heat balance in order to predict the temperature inside the furnace. The energy given by the electrodes is taken as an input, to which is added the energy obtained by chemical exothermic reactions that take place during the melting of scrap. By analyzing the system losses and the energy needed to melt that particular scrap mix, the temperature is derived. This can be summarized by the following equations:

Equation 2 defines the heat balance between input and output

\[
\text{Eq2.} \quad \text{Electric} \ E + \text{Chemical} \ E - \text{Losses} - Q = 0
\]

Where Q, the heat flowing to scrap, can be calculated as follows by equation 3

\[
\text{Eq3.} \quad Q = m[c_{p,scrap}(T_{melt} - T_{preheat}) + Lf_{scrap} + c_{p,melt}(T_c - T_{melt})]
\]

Then Tc is calculated by solving Eq3

\[
\text{Eq4.} \quad T_c = T_{melt} + \frac{Q}{c_{p,melt}} - \frac{1}{c_{p,melt}}[c_{p,scrap}(T_{preheat} + T_{melt}) - Lf_{scrap}]
\]
Where $c_p$ and $L_f$ are respectively the heat capacity and the latent heat of fusion of the scrap mix, $T_{melt}$ the temperature at which all scrap is molten, while $T_{preheat}$ is the temperature at which scrap exits the preheating shaft. Although it is not this project’s goal to analyze the model used, but instead to identify potential flaws by applying statistical analysis. Therefore, the model specifications and parameters were not shared by the company and consequently will not be published in this report. Despite this, it is crucial, in order to make educated guesses and decision on how to operate, being aware of the basic principles of the model used by the company.

2.5. **Topology and Topological Data Analysis**

Topology is the subfield of mathematics that concerns itself with the study of shapes [14]. Its origins can be tracked back to the 18th century with the publication of the paper “Seven bridges of Königsberg” by Leonhard Euler [15]. The Swiss mathematician asked himself if it was possible to cross the seven bridges in the city only once and return to the starting point. He constructed a network of nodes and edges and declared that it was in fact impossible. It is believed that this was the inspiration and the start of modern topology.

Topology has been studied for its own sake for the last 250 years but in the last decades it was found that it has many applications to real world problems. The one in which this project is interested in is the capability of analyzing complex multidimensional datasets, and this new area of study is called topological data analysis.

There are three main properties that are vastly used in topological data analysis that led it become a powerful tool in extracting knowledge from a dataset. Contrary to the Euclidian geometry, which is ruled by the concept of “congruence”, many transformations are allowed in topology before calling two objects different [14].

- **Coordinate invariance**: Topology measures shapes and properties that don’t change even with rotations or a change in the coordinate system. Thanks to coordinate invariance a circled is treated equal as an ellipse [16].

- **Deformation invariance**: Bending or stretching are allowed in topology; two objects remain equal until they are under continuous deformations. This property is commonly used by all human beings on a daily basis: it is possible in fact to recognize letter no
matter the used font. This is because letters share common feature that remain the same no matter how they are written. With the same concept in mind, it is also possible to realize that a donut has, from a topological point of view, the same shape as a coffee cup [15].

- **Compressed representation:** Continuity is often hard to treat so topology focuses also on connectivity as a way to represent continuous shapes. For instance, a sphere is a continuous space of points, but it is possible to represent it through an icosahedron, which instead is a finite element with 20 faces, 12 vertices and 30 edges. Figure 3 helps visualizing the approximation.

![Figure 3 Icosahedron inscribed in a cube and a sphere [17]](image)

## 2.6. **TDA using Ayasdi™**

The software used for this project named Ayasdi™, is able to analyze large array of data using TDA. It is based on two basic principles: “data has shape” and “shape has meaning” [18]. These two catch phrases hold all the philosophy of TDA, and describes easily what the logic behind this new approach is, compared to the more conventional ones. Consider an array of data with a finite number M of elements. These elements (represented by the symbol “E”) are described by a N number of properties (represented by the symbol “P”). This array then will look like a table
with M rows and N columns or vice versa as shown in Table 1, where $x_{i,j}$ represent the value of the property $P_j$ for the element $E_i$.

\[
\begin{array}{cccc}
P_1 & P_2 & \cdots & P_{N-1} & P_N \\
E_1 & x_{1,1} & \cdots & \cdots & x_{1,n} \\
E_2 & & & \cdots & \\
\vdots & & \ddots & \ddots & \vdots \\
E_{M-1} & & & \cdots & \\
E_M & x_{m,1} & \cdots & \cdots & x_{m,n} \\
\end{array}
\]

Table 1 Example of dataset

Instead of treating this dataset as a collection of numbers, TDA’s approach is visual. It assigns a shape to the dataset so further analysis can follow. It is counterintuitive to think that data has a shape although every element $E_i$ can be described as follows by equation 5:

\[
E_i = [P_1(i); P_2(i); \ldots; P_{N-1}(i); P_N(i)]
\]

Therefore, every element can be imagined as a point in a space of N+1 dimensions, and the all dataset as a cloud of points in that said space. The distribution of these points will define a shape of the cloud, and TDA analyzes it to extract the knowledge needed. Now instead of an unstructured mass of numbers it is possible to create a network from this cloud of points in order to be visualized in a two or three-dimensional space. In fact, representing an N-dimensional space is not an easy task and even if possible it is not easy to represent it visually. Ayasdi™
utilizes several algorithms called “Lenses” to collapse the N-dimensional space into a network that can be represented visually.

Figure 4 A circular cloud of point collapsed by a Lens [19]

Figure 4 represent a circular cloud of point collapsed by the lens “y-coordinate function” into a one-dimensional space. From here, it is easy to apply the same concept to a case where instead of starting from a two-dimensional space, there is N-dimensional one instead.

Figure 5 Division into overlapping sets [19]
The Cloud is then cut into overlapping sets, in a way that multiple points belong to more than one set as in Figure 5. The width of these sets is decided by a customizable parameter called “Resolution”. The higher the resolution the lower the width.

Points are then clustered by a measure of similarity called “Metric”. In this case the metric is the distance from one another, meaning that not every point will be clustered together but only the ones close to each other. That is the reason why in middle sets there are two clusters instead of one like in the ones on the top and bottom.

Two clusters that share one or more points are connected by an edge. The overlap of the slices is crucial to establish connections between clusters, and is regulated by another parameter called “Gain”. The higher the gain the bigger the overlap. In this way it is possible to represent a cloud of points into a two or three dimensional network which is possible to be visualized. As shown comparing Figure 4 and Figure 6, this method ensures that the original shape is approximated maintaining its important features, as the principles of topology dictate.

The network is then colored by an output variable than can be decided independently by the network construction and can lead to the discovery of the insights the analyst is looking for.
3. Method

3.1. Problem formulation

As already mentioned in the introduction, the project’s goal is to highlight possible flaws in the model that will be investigated in order to obtain an increase in accuracy. The company shared a dataset presenting 1138 charges with both records of measured temperature \( T_m \) and calculated temperature \( T_c \). 701 charges remained after filtering by eliminating the one with incomplete information or with value errors. Each charge is characterized by 104 parameters.

By plotting a graph of \( T_m \) against \( T_c \), it is possible to define the accuracy. Figure 7 represents the difference between the model and the temperature measurements. At each of them it was subtracted an average value for the tapping temperature. All the points are expected to lie close to dotted green line where \( T_m = T_c \) while instead almost 40% surpass the band of 100°C difference. A new parameter can then be defined named Delta Temperature (\( \Delta T \)). \( \Delta T \) is calculated as follows by equation 6:

\[
\text{Eq6.} \quad \Delta T = T_c - T_m
\]

This way it is possible to define three different categories of heats based on the value of \( \Delta T \):

- Over predicted charge (OPC): \( T_c > T_m, \Delta T \geq +30°C \)
- Under predicted charge (UPC): \( T_c < T_m, \Delta T \leq -50°C \)
- Accurately predicted charges (APC): \( T_c \approx T_m, -50°C < \Delta T < +30°C \)

The thresholds that define these three categories are arbitrary and were decided by the analyst. It is worth to notice the asymmetry on the zero of the APCs threshold. In fact, UPCs are more tolerated than OPCs. In the first case there is an overheat caused by the \( T_c \) being lower than the \( T_m \), and despite this is a substantial waste of energy and an increase in the wear of the refractory walls, the opposite instead can lead to more harmful consequences. In fact, OPCs may present some unmolten scrap caused by \( T_m \) being lower than \( T_c \). Solid scrap can clog the outlet when tapping the steel for further operations. Furthermore, the higher the amount is of unmolten scrap inside the furnace, the less material is at disposition during the rest of steel making operations. On top of that the composition of steel can be off target due to the fact that different kinds of
scrap melt at different temperatures due to different heat capacities. Therefore, there should be lower tolerance regarding how you defined a OPC. With these definitions in mind among the 701 charges, 23% are OPCs, 49% are UPCs and 28% APCs.

Model accuracy

\[ |T_c - T_m| = 100^\circ C \]

\[ T_c = T_m \]

Figure 7 Model accuracy


3.2. Estimation of gains

Before digging into the project’s methods, it is interesting to point out how much improving the predictive model could impact on the yearly economy of the steel plant. For practical reasons the calculations made are a rough estimation, and by no means want to be accurate, but at least they can provide an order of measure regarding how big the gains are once the model is improved.

Roughly the daily production of steel plus slag can be rounded to 2000 tons. By using the material properties published on Outokumpu’s website, the thermal capacity of steel is set to 460 J/kg°C [20]. As an approximation the thermal capacity of steel is used for all the material inserted in the furnace. The electric energy saved by improving the UPCs by 25°C can be calculated as follow:

\[ Q_{overheat} - Q_{accurate} = m c_p (T_{overheat} - T_{start}) - m c_p (T_{accurate} - T_{start}) = m c_p (T_{overheat} - T_{accurate}) = 2000 \text{ ton} \times 1000 \frac{\text{kg}}{\text{ton}} \times 460 \frac{\text{J}}{\text{kg}^\circ \text{C}} \times 25^\circ \text{C} = 23 \text{ GJ} \]

According to Eurostat.com, the cost of the electricity for industries in Sweden in 2015 was 0,059 €/kWh [21]. By approximating the energy saved only as electric energy this translates in

\[ 23 \times 10^9 \text{ J} \times \frac{1}{3.6 \times 10^6} \frac{\text{kWh}}{\text{J}} \times 0.059 \frac{\text{€}}{\text{kWh}} \approx 377 \text{ € per day} \]

The cost over the year becomes roughly 140000 €. Despite this number might appear exigous for a big company, the are several aspects that were not taken into consideration for this calculation. Most of the underpredicted charges far exceeds a 25°C ΔT, meaning even more energy can be spared. Also, even OPCs are a problem since the cost of reheating steel in the converter is much more expensive. Additionally, the presence of unmolten scrap can cause drastic changes to the final composition of steel, compromising the standard properties of the final product. On top of that, a more accurate temperature estimation reduces the wear of the furnace, nonetheless decrease the tap-to-tap cycle time. These are all savings that are hard to estimate but that can be much higher than the energy savings. Also environmentally speaking there is a big strive to reduce energy consumption. This translates into reduced carbon emissions, which is a big step to increase the sustainability of the steel making processes.
3.3. Database and network construction

Crucial to the analysis was the definition of the dataset used. The company provided a huge variety of data that needed to be recollected and filtered into one single dataset. As explained in the right part of Figure 1 Ayasdi™’s approach is data-first driven, meaning the more parameters are included in the analysis, the higher are the chances to find the right correlations in the data. Hence, it was not important to define what was relevant and what was not, even though a mild selection occurred. Although, it was crucial to transform the data in order to have comparable quantities among different heats. For this reason, most of the values were expressed in mass percentage. 104 parameters had been evaluated per heat and could be regrouped into several categories. Temperature was reported in three ways: Tc, Tm and ΔT. Metallic weight, slag weight per element and oxides mass percentage as well as the materials inserted by lancing operations were categorized as “Input”. The energy model was taken into account with all its different components (energy input, system losses and energy needed to melt the scrap mix etc.). On a later stage, nine parameters regarding what kind of materials were used in order to achieve the desired composition were added.

<table>
<thead>
<tr>
<th>Input / Network</th>
<th>Energy Model</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Weight</td>
<td>- Energy Input</td>
<td></td>
</tr>
<tr>
<td>- Composition</td>
<td>- Required Melting Energy</td>
<td></td>
</tr>
<tr>
<td>- Lance additives</td>
<td>- Energy</td>
<td></td>
</tr>
<tr>
<td>- Temperatures</td>
<td>- Energy Losses</td>
<td></td>
</tr>
<tr>
<td>- O₂</td>
<td>- Material Groups, Tested/Untested Materials</td>
<td></td>
</tr>
</tbody>
</table>

Figure 8 Graphical summary of the dataset

Once the database was assembled the software was able to start the analysis. It had been decided that only the category “Input” would be used for the network construction. In this way the energy
model, and its possible correlation with one or more physical parameters, could remain unbiased from the rest of the parameters. This could lead to stronger insights regarding the causes of the model inaccuracy. The dataset’s graphic summary is presented in Figure 8.

After the decision of which variables needed to be included into the network construction, metric resolution and gain were decided after a consultation between the author of this report and an expert in TDA analysis which provided the background knowledge needed to utilize the software. Lenses and metric regulate how the cloud of points, namely the dataset, is clustered in a two-dimensional space, while gain and resolution define respectively the amount of connections between clusters and their size. It was crucial to find, after several attempts, the right balance between the last two parameters in order to provide a meaningful shape to the data.

For reference the following parameters were applied to the “Input” column set to build the network:

- **Metric**: Variance Normalized Euclidean
- **Lens1**: PCA coord 1 (Res: 30 Gain: 5)
- **Lens2**: PCA coord 2 (Res: 30 Gain: 5)
- **Lens3**: ΔT (Res: 50 Gain: 5)
- **Lens4**: T_m (Res: 50 Gain: 5)

### 3.4. Hypothesis formulation

Once the data had been collected and analyzed, the factors that could lead to a wrong temperature estimation were investigated. The energy model, as explained in paragraph 2.4, is based on a mass and energy balance that calculates the temperature. It considers the energy input, the energy needed to melt the scrap mix and the energy losses. Therefore, it was possible to categorize several causes that would lead to a wrong temperature estimation into groups. Figure 9 summarizes them visually.
Since the energy given by the electrodes is well known, the mismatch between $T_c$ and $T_m$ can be either derived by a wrong estimation of the required energy to melt the scrap mix, or the energy losses. If the latter are calculated higher than reality, $T_c < T_m$ and vice versa. Regarding the former it is possible to identify two subcategories, systematic errors or a wrong material estimation. Systematic errors come into place when the energy model is using a wrong parameter or equation regarding one or more variables. A wrong material estimation affects both the mass of the material inserted in the furnace and the energy needed to melt them. Once this categorization was done it was possible to analyze the data to see where possible flaws may lie.
4. Results

4.1. The network

After have chosen $\Delta T$ as the output variable, namely the one ruling the color code, the results can be seen in Figure 10. Red, and in general warm colors, depict the highest values of $\Delta T$, representing the OPCs while cold colors stand for the negative values, therefore they represent UPCs. APCs are represented by yellow/light green colors. As mentioned in paragraph 2.6 the points shown are not single heats but rather clusters of them grouped together by a method of similarity. It is worth to specify it again because of the confusion that might be created by thinking the clusters are singular heats.
4.2. Confronting the color coding

The method used to analyze the network was to confront different color coding from different output variables to spot some similarities with $\Delta T$. Several parameters showed interesting behaviors if compared with the trend of $\Delta T$ and this led to a possible definition of the causes behind the mismatch between $T_c$ and $T_m$. First of all, it was necessary to define several groups in the original network to confront their behavior while changing color coding.

Figure 11 shows how the selection was made.
In Figure 12 APC1, APC2 and UPC1 show high values of oxygen percentage in steel, while OPC1 and APC3 have low values of it. High %O reflects a high volume of metallic oxides in the scrap mix.
Slag weight resembles the trend that can be spotted in Figure 12 where high values stay in the upper part of the network and the low values in the bottom one.
Figure 14 has a similar trend compared to Figure 12 and Figure 13. OPC1 and APC3 show low values of %Mo while the upper part of the network shows mostly high values. The distribution of values also is quite interesting. There are three bell curves, meaning there are probably three classes of steel. The one centered in the lower values has few outliers and many charges. The center one is quite small and not very spread, while the one situated in the upper part of the scale present many scattered values all regrouped in the last band but whit similar size of the center one.
All the groups present a high range of values between them. Both high and low values of energy losses can be found in either OPCs or UPCs. There seem to be no bias regarding one of the groups or part of it therefore it seems unbiased regarding $\Delta T$. 
While between APC1, APC2 and UPC1 the values of Energy demand look comparable, OPC1 presents lower values and mostly concentrated in the circled flare (named F1) in the upper right corner of the group. Differently from Figure 12 and Figure 13, where APC3 was following the same trend of its neighbor group, it presents higher values than OPC1, comparable to what can be found in the rest of the network. It is interesting to notice that the bottom tail of the values distribution presents a high number of clusters. This is due to the fact that the software recollects all the scattered clusters in the last band, meaning there are a higher number of outliers when energy demand is low than when is high.
This and future pictures will present a grey color that never occurred in previous cases. Grey means that the quantity chosen as outcome variable is not present in the clusters.

Figure 17 Turnings quality A

As for Figure 16, F1 in OPC1 present a considerable different set of values compared to the rest of its group and the overall network. In particular, almost all the clusters present 0% of turnings. It is also noticeable how that the high values lie in UPC1, APC1 and APC2 but not in APC3 which has instead a similar trend compared to OPC1.
In Figure 18 the values seem spread among groups with no particular trend aside for a part of F1 in OPC1 where a considerable amount of cluster presents a null percentage of turnings. The distribution also is quite peculiar: there are many outliers collected together in the upper tail. This means there are a lot of scattered values when there’s a high percentage of quality B turnings in the scrap mix while almost no outliers when the presence is low.
Reducing agent’s low values lie mostly in OPC1 and APC3 with also some zones where the value is 0. Worth to notice is the peak in the distribution for a particular set of low values that are mostly present in F1.
“Untested” is referred to the fact that these materials are yet to be tested by the company to understand their melting energy requirements. The parameter is binary (either it’s tested or untested) but since Ayasdi™ groups several heats into the same cluster, it happens that not all charges within a cluster have untested materials. Therefore, there is more than one color that express the presence of untested materials: either the cluster is red which means that all the charges present untested materials, or it is yellow which implies that only some charges has untested materials in them. Despite some isolate cases spread around the network, almost all the clusters that have a partial or total presence of untested material are concentrated in F1.
5. Discussion

Using Figure 9 as a reference, it is possible to categorize all the pictures collected in paragraph 4.2 and discuss them one by one to understand where the model’s flaws might lie.

*Energy Losses:* Figure 15 shows the trend of the energy losses. Confronting it with Figure 10 the overall distribution of values seems unbiased and does not provide any insights regarding a possible correlation with ∆T. If that would be the case in fact, it would have been possible to spot zones of high/low energy level concurrently with zones of high/low ∆T. Instead all the groups present pretty much the same distribution of values. Although, even if it is impossible to know if the system losses are calculated correctly, they are probably not a concurrent cause to a wrong temperature estimation. Worth to notice, the energy losses present a considerable higher number of outliers for high values rather than for low ones. In fact, the right tail is far exceeding in high than the left one, where the outliers of low values of energy losses lie.

*Systematic Errors:* Among all the parameters belonging to “Input”, Slag weight, %O and %Mo showed a similar trend that is worth to look into it. Slag weight and %O in steel are quantities connected with each other and it is understandable that they would follow a common trend. In fact, a high %O in steel means the scrap mix has a high presence of metal oxides in the form of Me_xO_y. In order to separate the metal from the oxygen, materials with high Oxygen affinity are added. The most common used are silicon and aluminum forming SiO_2 and Al_2O_3. Considering the definition of slag basicity expressed by Eq1, the more Silicon added the more the slag basicity decrease. An acid slag is detrimental to the refractories walls and to the refining qualities, therefore it must be avoided. To compensate this fact more CaO and MgO are added to the scrap mix, hence the high content of slag weight.

It is worth to mention though, that the quantities that can be found in “Input” are part of the column set that built the network, therefore it is possible that the trend shown is in reality a software’s artifact. In fact, Ayasdi™, by how the setup was chosen, clusters together steel that in starter N-dimensional space present similar physical characteristics, therefore it is reasonable to image that heats with similar characteristics such as slag weight, %O are in each other’s proximity.
Material Estimation: Figure 16 shows an interesting behavior regarding the energy demand. It represents the amount of energy needed to melt the scrap mix and can be calculated as follows.

\[ E_{d, \text{calc}} = mc_{p, \text{scrap}}(T_{\text{melt}} - T_{\text{preheat}}) + mL_{f, \text{scrap}} \]

It is noticeable that all the OPCs present the lowest amount of energy demand. This is in line with the idea that the energy demand is miscalculated, and that’s one of the main reasons that cause a wrong temperature estimation. In fact, if the calculated energy demand is lower than the real requirements, \( T_c > T_m \). This is easily demonstrated by combining Eq3 and Eq9.

\[ Q = E_{d, \text{calc}} + mc_{p, \text{melt}}(T_c - T_{\text{melt}}) \]

Which leads to a set of two equations

\[ Q = E_{d, \text{calc}} + mc_{p, \text{melt}}(T_c - T_{\text{melt}}) \]

\[ Q = E_{d, \text{real}} + mc_{p, \text{melt}}(T_m - T_{\text{melt}}) \]

Since \( Q, m, c_{p, \text{melt}}, \) and \( T_{\text{melt}} \) are the same in both equations is possible to merge Eq11 and Eq12

\[ E_{d, \text{real}} - E_{d, \text{calc}} = mc_{p, \text{melt}}(T_c - T_m) \]

Since the difference on the left side of the equation is a positive number by the hypothesis, also the difference on the right part is a positive number meaning the calculated temperature is higher than the measured one. This is due to the fact that the latent heat of fusion and/or the heat capacity are higher than expected and more energy is absorbed by the scrap mix before melting, reducing the amount of heat that can increase the molten steel temperature. Vice versa if \( E_{d, \text{calc}} > E_{d, \text{real}}, T_c < T_m \).

Regarding UPCs is harder to estimate how much a wrong energy demand estimation impacts on the temperature calculations. In fact, the same levels of energy demand appear for both UPCs and APCs. While F1 and in general the OPCs showed lower levels compared to the neighbor accurately predicted groups, UPCs have comparable values to APCs, meaning there is probably a concurrent cause that affects the temperature estimation.
5.1. Flare study

Despite that a global solution was not found, it was possible to provide a local explanation of a region in the OPCs’ area previously named F1. The region interested in this study is the flare circled in Figure 21, a collection of 29 charges that mostly hold the same steel code or at least are part of the same steel family. Several other charges being part of the same big family are evenly spread across the whole network. Therefore, it is believed that the ones clustered together might present a common problem. Using Figure 20 as a reference it is possible to correlate the wrong energy estimation with a high content of untested materials. As already mentioned untested materials had not gone through a melting test and their melting energy is only an approximation. As shown in Figure 8, the relationship between untested material and the flare’s
temperature wrong prediction is validated by how the analysis was conducted. In fact, Ayasdi™
clustered together similar heats without taking in consideration the parameters regarding the
materials used. Therefore, there is no correlation between the network and the material used and
the fact that a lot of charges are clustered together, while also presenting the same feature in
terms of material used, is a strong unbiased correlation between the materials and ΔT.

Also the low presence of quality A and B turnings might affect the temperature prediction. In
fact, turnings are well known in terms of composition because they are internal scrap deriving
from various steps of the production line. A low percentage of those increases the content of
scrap with unknown or with approximated composition scrap, which increases the uncertainty
regarding the energy demand.

Another worthwhile parameter is the one visible in Figure 19. Reducing agents are connected to
the slag weight and %O content so it is not surprising that they all have the same trend, but is
interesting to notice that somehow it follows the ΔT trend, in case further studies want to dig
deeper in the matter.
6. Conclusions

With the help of TDA and Ayasdi™ it was possible to examine the EAF predictive model’s fallacy, and understand which parameters were influencing the temperature estimation the most. Systematic errors were not found therefore it is believed that the core idea behind the equations and parameters used to model the process are unbiased. Although, the trends shown in Figure 12, Figure 13 and Figure 19 raised some suspicion about the presence of a bias. The trends seen are hard to connect with ΔT, although a high %O slag weight can be found for both UPCs and APCs, and the same applies for low %O slag weight being present for OPCs and APCs. It is peculiar though that there are no OPCs with high %O slag weight or any UPCs with a low number of them. This is not a strong claim especially considering the risk of being a software’s artifact, but it might be something to look at. In fact, since these parameters were part of the ones used to create the network is it possible that the heats were clustered together also because of similar %O or slag weight. Therefore, in this case it is crucial to be careful when analyzing their distribution.

It was found that a better understanding of the composition of the materials used in the scrap mix, and especially their melting energy, is key in order to reach an adequate level of accuracy. Proof of this fact can be found in the estimation of the energy demand. The required energy values found in the OPCs are the lowest in all the network: there is no APC that present the same values of energy demand while showing a tolerable ΔT. For this reason, it is believed that most of the problems that caused high absolute values of ΔT lie in the calculation of the energy demand. It was not possible to globally explain the correlation between energy demand and ΔT, but a local solution regarding a small flare of 29 charges was found. Most of these charges contained material that had yet to been tested, and that the energy demand approximation related to these materials led to a wrong temperature estimation. In particular, the energy demand level in the flare is among the lowest in the all network and considering that almost all the untested materials lie in the same flare, it is safe to assume that the energy demand of those material is underestimated.

There are several strategies that can be adopted to increase the model accuracy. Regarding the OPCs, the energy demand needs to be increased. Without knowing the energy model it is hard
to estimate by how much, but there are several solutions available. In fact, it is possible to artificially increase the energy demand of the flare F1 to the average value of APC3, a neighbor group which presents similar characteristics in terms of slag weight, metal weight and energy losses but with a lower ΔT. Its energy demand is 3900 kWh higher than F1. Once the energy demand is changed, it is possible to check how much the ΔT of the charges presenting untested material is changed accordingly. Additional tweaks should lead to a more accurate energy demand value even without a melting test. A less elaborate method is to manually lower $T_c$ of the charges that presents untested materials since most of them are OPCs. The flare average ΔT is roughly +125 °C, therefore manually decreasing the $T_c$ by 100°C should fix most of the OPCs without creating too many UPCs. Regarding UPCs, it is hard to provide a solution since a global cause was not found. Although the average ΔT is roughly -45°C, therefore once some of the OPCs are fixed through a better estimation of the energy demand, an overall manual compensation of $T_c$ by adding 45°C should help increasing the accuracy. Replicating Eq7, a global underestimation of 45°C translates approximately into 6-7 kWh per ton of steel of overheating.

Summarizing:

- TDA proved to be an efficient method to investigate which parameters were influencing the most the temperature’s wrong prediction.
- With this kind of analysis, the equations and parameters that build the model resulted unbiased.
- There is a visible trend that connects Slag weight and %O in steel to a wrong estimation of the temperature, but there is no strong proof that supports this claim. There is also the possibility that it can be a software’s artifact.
- Most of the OPCs are caused by a miscalculation of the energy needed to melt the scrap mix. An underestimation of the latter in fact, it is proven to cause $T_c > T_m$. Although an overestimation of the energy demand doesn’t explain fully the behavior of the UPCs, but it is believed to be a complement factor. Further analysis need to be done in order to unravel the causes behind a miscalculation of temperature in the UPC area.
7. Further work

Further work with new data is needed in order to confirm the claims presented by this project. Unfortunately, the dataset given by the company was not big enough to create a test group that could have been tested independently. Considering that during these months of work the company collected new data it should be easy to retrieve new charges to analyze. On a secondary level, if the claims hold up, it is probably worthwhile to change the dataset focusing more on the ingoing materials. Is it believed to be possible to find the more troublesome materials, but most of the insights regarding that cannot be shown with the current settings. A possible route into improving the analysis could be transforming the parameter regarding the tested/untested material. In fact, the parameter shown in Figure 20 is binary: either there is untested material or there is not. Changing this parameter from a binary system to mass percentage of untested material in the scrap mix could lead to much stronger claims regarding its influence in the temperature estimation.

Another viable route is categorizing OPCs and UPCs to create a manual offset to the model: by understanding what causes an under or over prediction of the temperature and by how much, it is possible to manually compensate the temperature given by the model without provide any change to the model itself. This method could be helpful to the current state correcting the temperature miscalculation, while the needed modification, which presumable will need some time, are applied to the model.
Bibliography


