



HAL
open science

Toward the implementation of mid-infrared spectroscopy along the processing chain to improve quality of the tomato based products

Sylvie Bureau, Alexandre Arbex de Castro Vilas Boas, Robert Giovinazzo,
Benoit Jaillais, David Page

► To cite this version:

Sylvie Bureau, Alexandre Arbex de Castro Vilas Boas, Robert Giovinazzo, Benoit Jaillais, David Page. Toward the implementation of mid-infrared spectroscopy along the processing chain to improve quality of the tomato based products. *LWT - Food Science and Technology*, 2020, 130, pp.109518. 10.1016/j.lwt.2020.109518 . hal-03111799

HAL Id: hal-03111799

<https://hal.inrae.fr/hal-03111799v1>

Submitted on 16 Jun 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

1 Toward the Implementation of Mid-Infrared Spectroscopy along the processing chain to improve quality
2 of the tomato based products

3

4 Sylvie Bureau¹, Alexandre Vilas-Boas¹, Robert Giovinazzo², Benoit Jaillais³, David Page¹

5

6 ¹ INRAE, Avignon Université, UMR408 Sécurité et Qualité des Produits d'Origine Végétale, 84914 Avignon,
7 France, F-84000 Avignon, France.

8 ² SONITO, Interprofession de la tomate destinée à la transformation. Maison de l'Agriculture - Site
9 Agroparc Bat. B - 84912 Avignon Cedex 9, France.

10 ³ Unité de Statistiques, Sensométrie, Chimiométrie, INRAE, ONIRIS, Rue de la Géraudière, CS 82225, 44322
11 Nantes Cedex 3, France.

12

13 * the corresponding author:

14 Sylvie Bureau, PhD

15 phone: +33 432722509

16 e-mail: sylvie.bureau@inrae.fr

17

18

19 Abstract

20 The mid-infrared spectroscopy (MIRS) was investigated as a tool to improve the quality of tomato products
21 considering its implementation at different steps along the processing chain.

22 Models have been developed using partial least square (PLS) regression to predict the quality of raw and
23 processed tomatoes. A relevant method (Multi-year Combining models) consisting in adding early-season
24 tomatoes data within models developed using data of previous years, was shown as the most efficient and
25 adapted to realistic industry conditions. MIRS predicted, in external validation, soluble solids content (R^2
26 0.95), titratable acidity (R^2 0.88) and dry matter content (R^2 0.81) with a high accuracy of 0.1°Brix, 2.8 mmol
27 H^+ /kg and 0.4% respectively.

28 Secondly, MIRS was used to classify tomato products depending on processing methods (hot- or cold-
29 break) or varieties using factorial discriminant analysis (FDA) based only on spectral data. MIRS was
30 assessed as an efficient tool to classify processed tomato purees according to process, year and variety,
31 more accurately than the classification obtained with the reference data.

32 A possible implementation of MIRS was suggested at three strategic steps along the processing chain to i)
33 characterize the incoming raw material, ii) monitor the matrix changes during processing and iii) control
34 the final products.

35

36 Keywords: Industry-type tomato, quality, ATR-FTIR, prediction, classification.

37

38 Highlights:

39 - Strategies are setup to build robust models to predict tomato quality traits.

40 - MIRS allows an accurate classification of hot-break and cold-break tomato products.

41 - An efficient implementation of MIRS along the processing chain is proposed.

42 Introduction

43 Processed tomato trade is a competitive market, with a few major producers such as California (11
44 Mt/year), Italy (6 Mt/year) or China (5 Mt/year), and a number of smaller ones, such as France producing
45 180 000 t/year, but also importing more than 100 000 t/year. Tomatoes are processed into various base
46 products such as raw juice, low concentrated 'passata', and up to highly concentrated tomato paste. Base
47 products are then used as ingredients to generate various manufactured products such as soups, ketchup
48 or sauces. To promote their products, most producers act on quality, leveraging on variety and local
49 production, but also by developing specialties from juices directly concentrated at the right expected dry
50 matter content. This avoids diluting highly concentrated tomato pastes as traditionally operated by many
51 industrial tomato users. This trend results in an increased demand from the companies to sort and pay the
52 raw material according to their quality. This includes not only the traditional soluble solids content (SSC
53 expressed in degree Brix), but also the real dry matter content (DMC, including also **insoluble** component,
54 more likely correlated to viscosity) and some new sorting criteria which should be developed. For example,
55 the ability to process a viscous and colored product, or the ability to determine when the product reaches
56 the expected quality according to guidelines during and after the manufacturing of products would be an
57 achievement. The implementation of infrared tools throughout the production chain is therefore an issue
58 for producers and processors to reach these objectives. This technique is already used in many other
59 productions regarding quality targets. In the dairy industry, mid-infrared analyzers (MIRS) are used since
60 1964 and improved over years to provide a rapid determination of fat, protein and lactose content of milk
61 (Barbano and Clark, 1989; Lynch et al., 2006). Today, MIRS assists most payment of milk and dairy
62 products. Concerning cereals, MIRS is used to classify flours according to landraces or technological
63 treatments (Cozzolino, 2014) and to determine their contents in proteins, lipids, ash and moisture (Sujka
64 et al., 2017; Shi and Yu, 2017) and even further the intestinal digestibility of their proteins (Shi et al., 2019).
65 Tomato industry still barely uses MIRS, despite strong needs for prediction tools associated to quality.
66 In the order of trade relevance, quality attributes of tomato products are their rheological properties
67 (determining whether they are more or less viscous), their color (preferred as deep red as possible), and
68 still to a lower extend, their taste and aroma. Viscosity mainly depends on processing methods. The critical
69 steps are the breaking temperature (i.e. temperature at which fruits are crushed and initially heated) and
70 the progressive juice concentration by thermal treatment under vacuum (Barrett et al., 1998; Page et al.,
71 2012). Some cultivars were also selected for their ability to produce various levels of viscosity (Svelander
72 et al., 2010; Ayvaz et al., 2016). However, the biochemical and physical factors driving puree viscosity
73 remain not fully understood, and therefore viscosity is still empirically controlled in industry. Relationships

74 between viscosity and microstructure (particle size and shape and serum viscosity), dry matter content
75 (DMC) or pectin composition have been established (Barrett et al. 1998; Anthon et al., 2002; Moelants et
76 al. 2014; Santiago et al. 2017), but no direct model taking into account those parameters allows an
77 accurate prediction of puree viscosity from those biochemical data. As soluble solids content (SSC) has
78 been partially correlated to DMC, the refraction index (which allows for a rapid evaluation of the SSC,
79 expressed in Brix degree) is currently used all along the production chain as an evaluation of DMC, being
80 often considered as an indirect indicator of the viscosity. Some companies are even using a price-increase
81 according to SSC to encourage the incoming of high SSC tomatoes in the factory, expecting these tomatoes
82 to also have a high DMC (Foolad, 2007). However, DMC corresponds not only to SSC (mainly sugars and
83 acids) but also to insoluble solids (such as pectins and other cell wall components, proteins, lipids,
84 pigments) and therefore DMC should be a more accurate marker of the rheological properties. On another
85 side, titratable acidity (TA) affects the taste in balance with sugars. Measuring SSC together with DMC and
86 TA is therefore relevant to follow tomato quality. But, as their measurement on fruit is time consuming,
87 SSC is generally the only measurement, and its correlation to other traits is empirically expected. However,
88 the relationships between SSC, dry matter content and puree viscosity become weak when a large
89 variability of genotypes and various growing conditions are taken into account (Arbex de Castro Vilas Boas
90 et al., 2017), and therefore, SSC is becoming of poor interest to predict DMC or viscosity. Color and taste
91 mainly controlled by sugar, acid, volatile and lycopene content, are all strongly dependent on genetic
92 factors as well as on the ripening stages at harvest (Saha et al., 2010; Figas et al., 2015). The processing
93 treatments also affect the biochemical composition of puree (Svelander et al., 2010; Wilkerson et al., 2013;
94 Lijima et al., 2016; Page et al., 2019). Still, neither global model nor easy-to-measure parameters are
95 available to predict or measure their real influence.

96 Using MIRS coupled with the Attenuated Total Reflectance (ATR), provides a solution well adapted to
97 aqueous samples such as juices and purees (Kemsley et al., 1996; Garrigues and Rambla, 1998). MIRS
98 allows an accurate evaluation of dry matter content (DMC), soluble solids content (SSC) and titratable
99 acidity (TA) based on one single spectrum acquired in a few seconds, compared to the time-consuming
100 reference methods (Beullens et al., 2006; Scibisz et al., 2011). Such an efficiency is compatible with the
101 cadency required for the grading of incoming raw tomatoes when trucks deliver them to factories. Quality
102 traits such as sugar content, pH and viscosity are also predicted in hot-break cooked tomato juices by MIRS
103 (Wilkerson et al., 2013; Ayvaz et al., 2016).

104

105 Despite those relevant results, strategies to adapt this technique to real industry conditions remain poorly
106 documented. Therefore, the objective of this paper was to test several options to implement MIRS at three
107 specific steps of the value chain of tomato products:

108 i) On fresh fruits, to develop accurate PLS models in order to predict, in a single run, a complete
109 composition of the raw materials before processing such as dry matter content (DMC), soluble
110 solids content (SSC) and titratable acidity (TA) instead of the only Brix degree actually
111 measured as a biochemical quality trait of tomato. Here, strategies were compared to gain in
112 efficiency to develop models taking into account industrial habits and constraints. (Figure 1).

113 ii) On fresh tomatoes and their corresponding purees, to verify if samples from an experimental
114 design including genotypes x years x processing conditions can be discriminated using
115 discriminant analysis (FDA). As the processed purees exhibit a large variability of quality traits,
116 the objective was to evaluate if MIRS could detect puree variability according to the
117 characteristics of the raw tomatoes.

118 iii) And on manufactured products, to assess the MIRS accuracy as a tool for assisting quality and
119 traceability control. This was performed using both, data of our laboratory and data from the
120 industry to measure whether correlations can be found between MIRS and quality
121 measurement currently achieved by industry.

122

123 1. Material and methods

124 1.1. Plant materials and processed samples

125 1.1.1. Fresh tomatoes

126 Tomatoes were harvested over two years (2014-2015), all over the production area in France. In 2014, 102
127 samples from 30 varieties were collected in the South-East (Vaucluse and Bouches-du-Rhône Counties) as
128 well as in the South-West (Lot-et-Garonne County) of France, from the 24th of July to the 10th of September
129 at breaker, ripe and overripe ripening stages. In 2015, 144 samples from 45 cultivars were collected in the
130 same areas, from the 20th of July to the 15th of September, but only at ripe and overripe stages. Samples
131 included a core collection of 14 genotypes, namely Caladou, Delfo, H1293, H1301, H1311, H9036, Impact,
132 Increase, ISI29714, JAG8810, Leader, Perfect Peel, Pietra Rossa and Terradou, which were planted every
133 year in every location to measure the inter-annual and the local variability.

134 For each sample, 15 fruits were randomly harvested on three plants, cut into pieces of around 2 cm³,
135 quickly frozen and stored at -20°C. Before analysis, the tomato pieces were thawed and homogenized in a
136 Waring blender. Fruit homogenates were used for the biochemical and spectral characterization.

137

138 1.1.2. Puree processing at the laboratory scale

139 In 2016 and 2017, four cultivars (Terradou, H1015, H1311 and Miceno) were cultivated in an experimental
140 design including two irrigation levels and two blocks per treatment (Arbex de Castro Vilas Boas et al.,
141 2017).

142 For each sample, about 1 kg of tomatoes was prepared as follow: a 1-cm slice was cut in the central part
143 of each fruit and slices were directly stored at -20°C, representing the fresh tomatoes. The rest of the fruits
144 were cut into 2-cm² pieces. All pieces were mixed and split into two similar samples dedicated to the hot
145 break (HB) or cold break (CB) standard processing (Page et al., 2012). Both processing routes used the
146 same heating and grinding energy, and only the order of each unit operation changed. Tomatoes for HB
147 purees were first heated (microwave oven, 900 w, full power, 0.9 sec/g of tomato) and then grinded (30
148 seconds in a Waring blender) whereas the CB tomatoes were first grinded, macerated at room
149 temperature for 30 seconds (allowing for intrinsic enzyme reactions) and then heated.

150 After cooking, purees were stored into 400-ml glass jars, pasteurized (100°C, 15 min) and stored at 4 °C
151 until analyses. A total of 336 samples were characterized in 2016 and 2017, as fresh tomatoes, HB and CB
152 cooked purees.

153

154 1.1.3. Industrial products

155 In 2015, 140 tomato-based products (juices, purees and pastes) were collected from two factories located
156 in South-East (Tarascon) and South-West (Bergerac) of France. Their soluble solids content (SSC) and
157 viscosity were measured in parallel by the quality control of the factories and by our laboratory.

158

159 1.2. Reference analyses

160 Soluble solids content (SSC) was determined with a digital refractometer (PR-101 ATAGO, Norfolk, VA) and
161 expressed in °Brix at 20°C. Titratable acidity (TA) was determined by titration up to pH 8.1 with 0.1N NaOH
162 and expressed in mmol H⁺/kg of fresh weight using an autotitrator (Methrom, Herisau, Switzerland). The
163 dry matter content (DMC) was determined by weighing and drying 3 g of samples in air oven at 70°C to
164 reach a constant weight. The viscosity was measured as described by Arbex de Castro Vilas Boas et al.
165 (2017) using a viscosimeter (Anton Paar MCR 301, Graz, Austria). For the industrial products, consistency
166 was measured using a Bostwick consistometer (CSC Scientific Company, Fairfaix, USA) and according to
167 manufacturer's guidebook, results were expressed as arbitrary Bostwick unit (Bw). The lower the Bostwick
168 value, the higher the consistency.

169
170 1.3. Mid-Infrared Spectroscopy analyses
171 Spectra were recorded as described by Bureau et al. (2009) at room temperature with a Tensor 27
172 spectrometer (Bruker Optics, Wissembourg, France) equipped with a horizontal attenuated total
173 reflectance (ATR) sampling accessory composed of a zinc selenide (ZnSe) crystal with six internal
174 reflections and with a deuterated triglycine sulfate (DTGS) detector. Spectra were acquired between 4000-
175 650 cm^{-1} , with scanner velocity of 10 KHz, a background of 32 scans, and a resolution of 4 cm^{-1} . The
176 reference spectra were recorded using a blank ATR crystal every twenty samples. Between measurements,
177 the crystal was carefully cleaned using distilled water and dried with filter paper. In the range between
178 4000 and 400 cm^{-1} light penetrates from about 0.4 to 4 μm (Bureau et al., 2019). The total optical path is
179 therefore 2.4 μm at 4000 cm^{-1} and 24 μm at 400 cm^{-1} taking into account the six internal reflections.

180 1.4. Chemometrics

181 Spectral preprocessing and multivariate data analysis were performed as described by Bureau et al. (2013)
182 with Matlab 7.5 (Mathworks Inc. Natick, MA) software using SAISIR package (Cordella & Bertrand, 2014).
183 The absorption band around 2400 cm^{-1} , due to carbon dioxide, was discarded. Spectra were systematically
184 pretreated with the standard normal variate correction (SNV).

185 1.4.1. PLS modelling

186 Models were developed by partial least squares (PLS) regression on the fresh tomatoes harvested in 2014
187 and 2015 (see § 1.1.1). In PLS, orthogonal latent variables are iteratively constructed by maximizing the
188 covariance between the two matrices of data set, the spectral data (X) and the quality traits (Y, reference
189 data) (Nicolaï et al., 2007). In a first step, models were calibrated and validated by randomly splitting the
190 data set into a sub-set of calibration data (2/3 of the data) which was used to build the model, and a sub-
191 set of validation data (1/3 of the data) for which the content was predicted by using the previous built
192 model. The root mean square error (RMSE) between predicted and measured values was estimated to
193 evaluate the accuracy of the prediction. The random selection of calibration/validation data was repeated
194 10 times for each quality trait and the RMSE value was recalculated in order to examine the stability of the
195 model. In a second step, models were evaluated by an external validation, consisting in predicting the
196 composition of an independent validation data set, not used for the internal validation.

197 The performance of models was evaluated by the determination coefficient of calibration and of validation
198 (R_c^2 and R_v^2), determination coefficient of external validation (i.e. prediction) (R_p^2), root-mean-square error
199 of calibration and of validation (RMSE_c and RMSE_v) and root-mean square error of external validation
200 (RMSE_p). Finally, the ratio of prediction to deviation (RPD) corresponding to the ratio of the standard

201 deviation of the reference data to the RMSE was calculated. A RPD between 1.5 and 2 concerns a low
202 performance model which can only discriminate low from high values; a value between 2 and 2.5 indicates
203 a coarse quantitative prediction, and a value between 2.5 and 3 or above corresponds to good and
204 excellent prediction accuracy, respectively (Nicolai et al., 2007).

205 Three strategies were tested on the raw materials. The first one consisted in building models only based
206 on one-year data in 2014 and 2015 (named YPY), the second one in building one global model combining
207 total data of the two years (named GIC) and the third one in combining data of 2014 and a part of 2015
208 data corresponding to samples harvested during the early season of 2015 (before August, the 18th)
209 (named MYC) (Figure 1). All models were compared using internal and external validations when possible.

210

211 1.4.2. Discriminant analysis

212 Factorial discriminant analysis (FDA) was performed to test the ability of MIRS to discriminate samples
213 according to the known qualitative groups (genotypes, years and cooking procedures). FDA (Factorial
214 Discriminant Analysis) was performed on samples characterized in 2016 and 2017, as fresh, HB or CB
215 processed purees as described in § 1.1.2. It was carried out in two steps: 1) Principal Component Analysis
216 (PCA) was calculated on the spectral data to visualize the samples distribution according to the most
217 discriminating spectral ranges identified with the eigenvectors and 2) FDA was applied on the gravity
218 centers of each qualitative group assessed on the normalized principal component scores (Bertrand et al.,
219 1990).

220

221 2. Results and discussion

222

223 2.1. PLS models to predict quality traits of fresh tomatoes

224

225 2.1.1. Variability of the samples used to build models

226 To make our models as generic as possible, the sampling included 59 varieties harvested out of two regions
227 of France, over two years and at three maturity stages (see § 1.1.1). The values ranged from 3.6 to 7.5°Brix
228 for SSC, from 4.7 to 11.1% for DMC and from 30.2 to 81.7 mmol H⁺/kg for TA (Table 1). Fruit quality varied
229 for the reference data, and especially, a year effect was obvious on the relationship between SSC and DMC.
230 The groups of points of each year, 2014 and 2015, are parallel indicating that the classification of varieties
231 remained similar but, for a similar SSC, DMC revealed variations from 1 to 2% (Figure 2A). This was also
232 the case for the core of 14 varieties present in the 2014 and 2015 data sets. The year effect was not so

233 clear for the spectral data. On PCA plot calculated using spectra of the two years, the 2015 samples covered
234 a larger variability than the 2014 ones. This was probably related to the higher number of genotypes in
235 2015 than in 2014 (respectively 45 and 30) (Figure 2B).

236 Our results were in accordance with data already reported for processing tomatoes grown in California
237 counties between 2010 and 2014, and particularly for SSC which ranges from 3.2 to 7.2°Brix in (Wilkerson
238 et al., 2013; Ayvaz et al., 2016). The set of samples covered a large range of the variability generally
239 observed for industry-type tomatoes. This permits a standard robustness of calibration models, as
240 robustness is directly related to the variability of the samples (Nicolai et al., 2007).

241

242 2.1.2. Comparison of strategies to build accurate and robust models

243

244 a) Model calibration and validation

245 Any of the three strategies (YPY, GIC or MYC) gave accurate results to predict SSC. Similar results were
246 obtained for calibration and validation with R_c^2 and R_v^2 between 0.93 and 0.97 and $RMSE_{c \text{ and } v}$ between
247 0.13 and 0.16°Brix, leading to a RPD equal or above to 3.6 corresponding to a highly accurate prediction.

248 The MYC model obtained the highest RPD. Globally, RPD of models for SSC exhibited the highest RPD
249 among the quality traits (Table 2). Our results were similar to those previously obtained on tomato fruits

250 for the fresh market exhibiting SSC from 3.2 to 8.8 °Brix while our data exhibited no SSC above 7.5 °Brix
251 (Scibisz et al., 2011). Similar results were obtained on industry-type tomatoes grown in California, as SSC

252 is predicted with R_v^2 varying from 0.86 to 0.98 depending on the years, regions and varieties (Wilkerson et
253 al., 2013; Ayvaz et al., 2016). Our data confirmed that SSC is extremely well predicted by MIRS.

254 For the DMC parameter, models still exhibited high performance although the values were not as good as

255 for the SSC models. R_v^2 varied from 0.82 to 0.94 and RMSE from 0.24 to 0.49%. The RPD was in all cases
256 higher than 2.5, and then within the highly accurate values for predicting models (Nicolai et al, 2007). The

257 YPY models exhibited $RMSE_v$ of 0.25% in 2014 and 0.41% in 2015. The higher variability of fruit DMC in
258 2015 than in 2014 could explain the results. The differences between min and max values were 5.4% in

259 2015 but only 3.1% in 2014 (Table 1). RPD values indicated some differences of model performances. The

260 MYC model exhibited the highest values of R_v^2 (0.94) and RPD (3.9). Similar results with RPD of 4.8 are

261 obtained on tomatoes for fresh market using a dataset including a large number of traditional varieties
262 conferring a variability similar to that of our experiment (Scibisz et al., 2011).

263 As for DMC, models predicting TA did not present the same accuracy over the two years (Table 2). In 2015,

264 models exhibited higher R_v^2 and RPD than in 2014, even if the $RMSE_v$ remained close around 2.2 mmol

265 H⁺/kg. As for DMC, the range of TA was larger in 2015 (30.2-81.7 mmol H⁺/kg) than in 2014 (45.3-76.8
266 mmol H⁺/kg) (Table 1). This impacted the RPD values. However the prediction of TA remained within the
267 excellent RPD values (≥ 2.5). For TA, our results were similar to those obtained on fresh tomatoes by Scibisz
268 et al. (2011) and on industry-type tomatoes by Wilkerson et al. (2013).

269 So, combining data of different years in GIC models did not affect the model performance, except an
270 increase of the RMSE_v for TA, in comparison with the YPY strategy (Table 2). However, RPD remained
271 acceptable for the three predicted quality traits with values ≥ 2.5 . An interesting result came from the MYC
272 strategy. By introducing new data every year, and especially the data of the early tomatoes, the MYC
273 models were as accurate as the GIC models for SSC and DMC, despite less samples used to calculate the
274 models. For TA, the MYC and GIC models did not much differ for their R_v² but RMSE_v of MYC was the
275 highest, giving a RPD of 2.2. The gain of the MYC strategy was not obvious on the validation results for TA.

276

277 b) External validation of models

278 The external validation constitutes the ultimate validation of predicting models as samples used for
279 validation must differ from samples use for calibration belonging to another sample sets, and here to
280 another year. In this case, the YPY and MYC strategies exhibited contrasted results (Table 3). Concerning
281 the YPY models, predicting 2015 data with the 2014 models resulted in low R_p² and RPD and high RMSE_p
282 for the three quality traits, SSC, DMC and TA (Table 3). Predicting 2014 data with the 2015 models led to a
283 better prediction of SSC (RPD of 2.7) but not for DMC and TA. For DMC, RPD was 0.1 due to the RMSE_p of
284 2.54%, i.e. 5 times higher than the RMSE_v (Tables 2 and 3). For TA, RPD was 0.3 in relation with the RMSE_p
285 of 53.44 mmolH⁺/kg, i.e. 10 times higher than the RMSE_v (Tables 2 and 3). These results can be explained
286 by the difference of the fruit variability observed in the two years (Figure 2). The linear relationship
287 between the quality traits, SSC and DMC, may be maintained but contents of DMC changed for a same SSC
288 from one year to another.

289 The combination of data of several years significantly improved the models. The MYC models, which
290 combined all data of the first year and data of the earliest tomatoes of the second year (2014 + early 2015
291 until August, 18th) accurately predicted SSC, DMC and TA of the late tomatoes of 2015 (from August, 18th)
292 with similar RPD values (respectively 4.3, 2.8, 2.1) than those previously obtained (Tables 2 and 3). Adding
293 the early data of 2015 within the 2014 data (MYC models) led to a more efficient prediction of the late
294 tomatoes of 2015. Improving models by accumulating new data each year is an approach described by
295 Thomas and Ge (2000) as a passive approach consisting in acquiring calibration data over a sufficiently
296 long period. It tends to cover gradually the fruit variability by including variability such as seasons or years,

297 varieties, orchards in the calibration data to improve the model accuracy as already suggested (Peiris et
298 al., 1998; Peirs et al., 2003; Golic and Walsh, 2006; Bobelyn et al., 2010). Such approach is particularly
299 relevant for the every-day work of the tomato processors. The earliest tomatoes may be used to calibrate
300 and update models each year. The calibrated model can then be used for the rest of the season to
301 accurately predict the quality of the incoming production. At the end of the season, models can be
302 efficiently completed by the addition in the calibration of the most contrasted samples harvested during
303 the running year. They can be identified using their infrared signature in comparison with those already
304 placed on the cartography representing the tomato diversity, and only those samples can be analyzed by
305 reference methods. This method is a way to minimize the quantity of analyses to the most relevant ones,
306 and year after year, this approach leads to a progressive improvement of global models, by taking into
307 account variability of early and late tomatoes as illustrated in this paper (Figure 3).

308 This demonstration was focused on building PLS models for predicting the quality of raw tomatoes. But
309 one can assume that the same approach could be developed for the prediction of processed product
310 quality as shown by Wilkerson et al. (2013) and Ayvaz et al. (2016). In this case, including variability due to
311 the processing conditions should be considered in addition to all the other sources of variability.

312
313 2.2. Towards using MIRS to discriminate fresh fruits and processed products according to varieties, years
314 and processing conditions.

315 Discriminant analysis only based on spectral data was performed on a set of samples issuing from an
316 experimental design to evaluate the ability of MIRS to classify samples according to factors of interest,
317 such as varieties, years and processing conditions. The experimental design included four varieties, two
318 irrigation levels and was reproduced in 2016 and 2017 (see § 1.1.2). Samples were analyzed as fresh fruits
319 and after a hot or a cold break processing, and all samples were evaluated for their SSC, TA and DMC using
320 reference methods. Data exhibited a significant impact of the genotype ($F=70$, $\text{Prob}>5.3.10^{-34}$), water
321 scarcity ($F=70$, $\text{Prob}>2.8.10^{-15}$) and processing ($F=18.2$, $\text{Prob}>3.7.10^{-8}$) but no significant impact of the year.
322 However, the year affected standard deviations. In 2016, TA exhibited variations from 59.9 to 89.4 mmol
323 H^+ /kg FW while it ranged from 46.6 to 96.9 mmol H^+ /kg FW in 2017. The same trends were also observed
324 for DMC and SSC.

325 Factorial Discriminant Analysis (FDA) performed on spectral data classified the samples in their right
326 classes with only few confusion concerning the processing. All the 144 fresh samples were well classified,
327 89 among the 96 HB and 94 among the 96 CB were well classified giving a performance of classification of
328 100% for fresh, 93% for HB and 98% for CB (Table 4). The results on genotype classification was more

329 confused: for the two most contrasted genotypes, most samples were identified in their right classes (70%
330 for H1311 and 90% for Terradou), but the classification was less accurate for the two other genotypes, as
331 only 54% and 67% were well classified for H10 (H1015) and MIC (Miceno) respectively (Table 4). The
332 classification according to varieties was good when FDA was performed separately on each year. On fresh
333 tomatoes, for example, samples of Terradou in 2016 and H1311 in 2017 were 100% well classified whereas
334 for the other varieties the classification was at least higher than 88%.

335 Nevertheless, when considering fresh and processed samples separately, the FDA gave accurate
336 classifications of the genotypes. Each appeared as distinct and non-overlapped ellipses on the factorial
337 maps (Figures 4A and 4B). Moreover, the classification was partially reproducible from one year to the
338 other. When 2017 data were projected as illustrative data on the factorial map calculated with 2016 data,
339 ellipses from 2017 data remained distinct from one genotype to the others. For two of the genotypes
340 (H1311 and Terradou), 2016 and 2017 ellipses were in a very close area of the factorial map. The same
341 trend was observed in the reverse situation. On this FDA space, the distances between ellipses of each
342 year were greater for Miceno and H1015, but remained in the same region of the map, for fresh products
343 as well as for processed ones (Figure 4). The most significant spectral area distinguishing varieties was
344 between 1200 and 900 cm^{-1} corresponding to absorptions due to stretching and bending vibration modes
345 of sugars (Talari et al., 2017).

346 Altogether, discriminations based only on spectral data indicated that MIR was a powerful tool to follow
347 tomato quality during processing as it allowed a strict and accurate distinction of fresh, cold or hot break
348 samples. However, the infrared sensors exhibited some limits for distinguishing samples according to the
349 varieties when processed and fresh samples were considered altogether, and especially for those
350 exhibiting similar qualities. This last result should be challenged to a larger range of varieties, as our set of
351 data only contained four varieties, and to a high processing impact according to the genotype. Previous
352 studies on fresh fruits indicated that accurate genotype discrimination is made possible over a larger set
353 of varieties (Ibanez et al., 2019). To our knowledge, our studies was the first on tomato showing that the
354 same kind of distinction remained after fruit processing.

355
356 2.3. Toward the use of MIR tools for quality control of manufactured products

357 The products exhibited a diversity including purees, sauces and pastes giving a large variation of quality
358 traits. SSC varied from 5.3 to 36.5°Brix, pH from 3.9 to 4.7, TA from 44 to 319 $\text{mmol H}^+/\text{kg FW}$, DMC from
359 7 to 45 % and viscosity from 0 to 8 Bw unit (Table 5). SSC vs TA, SSC vs DMC and TA vs DMC exhibited
360 correlations with determination coefficient R_v^2 higher than 0.95. SSC was measured both in the Lab and in

361 the plants giving as expected similar values ($R_v^2 = 0.99$). TA and DMC were only measured in our laboratory,
362 pH and viscosity only in factories.

363 A first set of PLS models were built to measure their efficiency for predicting the product composition
364 taking into account their entire variability, from juice to paste (Table 5). SSC was extremely well predicted
365 by MIRS with R_v^2 of 0.99 and error ranged between 0.73 and 0.98 °Brix in Laboratory and in the factories
366 data respectively (Table 5). The RPD, higher than 12, confirmed that SSC can be predicted with a very weak
367 error using MIRS. TA and DMC measured in the Lab exhibited similar levels of prediction with R_v^2 higher
368 than 0.98 and RPD higher than 7.6, as we previously obtained in the other experimental assays. These
369 results were also in accordance with the strong internal correlation measured between those traits in this
370 set of samples. On the contrary, pH measured in the factories was predicted with a R_v^2 of only 0.51 and a
371 RPD of 1.4 (Table 5). The quality of our prediction was lower than that already obtained on industry tomato
372 (Ayvaz et al. 2016). This can be due to the lower size of our sample set (76 instead of 249 for the calculation
373 of the model), and its lower variability (pH ranged from 3.98 to 4.6 instead of 3.8 to 4.6). The prediction of
374 the viscosity (Bw) exhibited an apparent high accuracy. The R_v^2 of validation was 0.77 and the RPD 4.6.
375 However, the high contrast of viscosity between juice and paste and the low quantity of intermediate
376 samples were a concern regarding the statistical analysis. Therefore, in a second step, the models were
377 built after removing pastes in order to have a more continuous and homogeneous set of samples. For all
378 quality traits, RPD values decreased to values close to those obtained in our models on raw fruits (Table
379 5), assessing the accuracy of prediction on manufactured products.

380 With this restriction to juices and purees, models exhibited accuracy close to the models obtained by Ayvaz
381 et al (2016), which were also dedicated to tomato juices and purees (between 11 and 25 Bw) and
382 calculated with a large number of samples. Altogether, our results and those of Ayvaz et al. (2016)
383 indicated that predicting consistency by MIRS was certainly possible but hardly in actual realistic industry
384 conditions. Progress should be made in two directions. First, more universal and accurate measurements
385 of the rheological properties should be used as comparing Bostwick values of contrasted products such as
386 juices and purees should include specific corrections (Perona, 2005). Second, the rheology of tomato
387 products does not enforce the same mechanical properties depending on their concentrations. Those
388 properties depend upon biochemical and physical characteristics such as pectin dissolution and
389 modification, particle size and shape and particle packing (Bayod et al., 2008). Each of those characteristics
390 may have diverse MIRS signature, and this could explain why PLS models including pastes and less
391 concentrated products gave less accurate results. For the prediction of consistency, models per classes of
392 products, using a large variability within each class, should lead to models more accurate and adapted to

393 the real industry activity. Combination of models may also rise to accurate results to predict intermediate
394 products between purees and pastes, but require a more specific study.

395

396 Conclusions

397 Our results confirmed that MIRS is a powerful decision-making tool to assist the industry for the
398 improvement of the quality of tomato-based products all along the processing chain. In the realistic
399 industrial context, we demonstrated that MIRS could enhance industrial management at three strategical
400 steps:

- 401 - For the incoming tomatoes: as, in a single measure, not only SSC but also TA and DMC can be
402 predicted. MIRS gives a new framework for grading tomatoes regarding their quality.
- 403 - For assessing the processing: as Fresh, HB and CB samples can be discriminated, and considering
404 that the sorting of fresh or processed samples was accurate, this indicates that MIRS is a powerful
405 sensor to improve the product traceability before, during and after the processing step. Therefore,
406 development of databases of MIRS spectra needs to be achieved in a large industrial context.
- 407 - For the post-processing quality and trade management, as most quality traits of manufactured
408 products (SSC, DMC, acidity, viscosity) seemed to be predictable, MIRS could help for a more
409 pragmatic and complete verification of the accordance of manufactured products regarding the
410 specification books.

411 MIRS signature is easy and rapid to acquire on homogenous samples such as purees, liquids and pastes
412 compared to the classical measurements by reference methods. MIRS coupled with chemometrics greatly
413 increases the possibilities to enhance the quality, by a better management of raw materials and processed
414 products at all steps of the production chain. Our results give strategies for an industrial development,
415 including the accumulation of MIR data over years to integrate in calibration, and to gradually improve the
416 accuracy and the robustness of prediction.

417

418 Acknowledgment:

419 Alexandre Vilas-Boas was supported by a grant from CAPES and from the Brazilian Ministry of Education.
420 Authors gratefully acknowledge the two French companies (Provence Tomates S.A., Tarascon, France, and
421 Tomates d'Aquitaine S.A.S., Bergerac, France) for delivering tomato products and their corresponding
422 characteristics. Authors thank Patrice Reling, Caroline Garcia and Marielle Bogé for their technical support.

423

424 This research did not receive any specific grant from funding agencies in the public, commercial, or not-
425 for-profit sectors.

426 References

- 427 Anthon, G.E., Sekine, Y., Watanabe, N., & Barrett, D.M. (2002). Thermal Inactivation of Pectin
428 Methylesterase, Polygalacturonase, and Peroxidase in Tomato Juice. *Journal of Agricultural and*
429 *Food Chemistry*, 50, 6153-6159.
- 430 Arbex de Castro Vilas Boas A., David, D., Giovinazzo, R., Bertin, N., & Fanciullino, A. L. (2017). Combined
431 Effects of Irrigation Regime, Genotype, and Harvest Stage Determine Tomato Fruit Quality and
432 Aptitude for Processing into Puree. *Frontiers in Plant Science*, 8, 725.
- 433 Ayvaz, H., Sierra-Cadavid A., Aykas, D., Mulqueeny, B., Sullivan, S., & Rodriguez-Saona, L. (2016).
434 Monitoring multicomponent quality traits in tomato juice using portable mid-infrared (MIR)
435 spectroscopy and multivariate analysis. *Food Control*, 66, 79-86.
- 436 Barbano D.M., & Clark J.L. (1989). Infrared milk analysis – Challenge for the future. *Journal of Dairy Science*,
437 72, 1627-1636.
- 438 Barrett, D., Garcia, E., Wayne, J. (1998) Textural modification of processing tomatoes. *Critical Reviews in*
439 *Food Science and Nutrition*, 38(3), 173-258.
- 440 Bayod, E., Pilman Willers, E., Tornberg, A. (2008). Rheological and structural characterization of tomato
441 paste and its influence on the quality of ketchup. *LWT - Food Science and Technology*, 41, 1289-
442 1300.
- 443 Bertrand, D., Courcoux, P., Autran, J.C., Meritan, R., & Robert, P. (1990). Stepwise canonical discriminant
444 analysis of continuous digitalized signals: application to chromatograms of wheat proteins. *Journal*
445 *of chemometrics*, 4, 413-427.
- 446 Beullens, K., Kirsanov, D., Irudayaraj, J., Rudnitskaya, A., Legin, A., Nicolai, B. M., & Lammertyn, J. D. (2006).
447 The electronic tongue and ATR-FTIR for rapid detection of sugars and acids in tomatoes. *Sensors and*
448 *Actuators B-Chemical*, 116(1-2), 107-115.
- 449 Bobelyn, E., Serban, A. S., Nicu, M., & Lammertyn, J., Nicolai, B. M., Saeys, W. (2010). Postharvest quality
450 of apple predicted by NIR-spectroscopy: Study of the effect of biological variability on spectra and
451 model performance. *Postharvest Biology and Technology*, 55(3), 133-143.
- 452 Bureau, S., D. Ruiz, D., Reich, M., Gouble, B., Bertrand, D., Audergon, J. M., & Renard, C.M.G.C. (2009).
453 Application of ATR-FTIR for a rapid and simultaneous determination of sugars and organic acids in
454 apricot fruit. *Food Chemistry*, 115(3), 1133-1140.
- 455 Bureau, S., Quilot-Turion, B., Signoret, V., Renaud, C., Maucourt, M., Bancel, D., & Renard, C.M.G.C. (2013).
456 Determination of the Composition in Sugars and Organic Acids in Peach Using Mid Infrared

457 Spectroscopy: Comparison of Prediction Results According to Data Sets and Different Reference
458 Methods. *Analytical Chemistry*, 85(23), 11312-11318.

459 Bureau, S., Cozzolino, D., & Clark, C.J. (2019). Contributions of Fourier-transform mid infrared (FT-MIR)
460 spectroscopy to the study of fresh and processed fruits and vegetables: A review. *Postharvest
461 Biology and Technology* 148, 1-14.

462 Cordella, C. B. Y., & Bertrand D. (2014). SAISIR: A new general chemometric toolbox. *Trends in Analytical
463 Chemistry*, 54, 75-82.

464 Cozzolino D. (2014). An overview of the use of infrared spectroscopy and chemometrics in authenticity
465 and traceability of cereals. *Food Research International*, 60, 262-265.

466 Figas, M.R., Prohens, J., Raigon, M.D., Fita, A., Garcia-Martinez, M.D., Casanova, C., Borrás, D., Plazas, M.,
467 Andujar, I., & Soler, S. (2015). Characterization of composition traits related to organoleptic and
468 functional quality for the differentiation, selection and enhancement of local varieties of tomato
469 from different cultivar groups. *Food Chemistry*, 187, 517-524.

470 Foolad M. (2007). Genome Mapping and Molecular Breeding of Tomato. *International Journal of Plant
471 Genomics*. 52 p.

472 Garrigues, S., & Rambla, F. J. (1998). Comparative study of reflectance cells for PLS-FTIR determination of
473 sugars in soft drinks. *Fresenius Journal of Analytical Chemistry*, 362(1), 137-140.

474 Golic, M., & Walsh, K. B. (2006). Robustness of calibration models based on near infrared spectroscopy for
475 the in-line grading of stonefruit for total soluble solids content. *Analytica Chimica Acta*, 555(2), 286-
476 291.

477 Ibáñez, G., Valcárcel, M., Cebolla-Cornejo, J., Roselló, S. 2019. FT-MIR determination of taste-related
478 compounds in tomato: a high throughput phenotyping analysis for selection programs. *Journal of
479 the Science of Food and Agriculture*, 99, 5140–5148.

480 Kemsley, E. K., Holland, J. K., Defernez, M., & Wilson, R. H. (1996). Detection of adulteration of raspberry
481 purees using infrared spectroscopy and chemometrics. *Journal of Agricultural and Food Chemistry*,
482 44(12), 3864-3870.

483 Iijima, Y., Iwasaki, Y., Otagiri, Y., Tsugawa, H., Sato, T., Otomo, H., Sekine, Y., & Obata, A. (2016). Flavor
484 characteristics of the juices from fresh market tomatoes differentiated from those from processing
485 tomatoes by combined analysis of volatile profiles with sensory evaluation. *Bioscience,
486 Biotechnology, and Biochemistry*, 80, 2401–2411.

487 Lynch, J. M., Barbano, D. M., Schweisthal, M., & Fleming, J.R. (2006). Precalibration Evaluation Procedures
488 for Mid-Infrared Milk Analyzers. *Journal of Dairy Science*, 89, 2761–2774.

489 Moelants, K.R.N., Cardinaels, R., Van Buggenhout, S., Van Loey, A.M., Moldenaers, P. & Hendrickx, M.E.
490 (2014). A Review on the relationships between processing, food structure, and rheological
491 properties of plant-tissue-based food suspensions. *Comprehensive Reviews in Food Science and*
492 *Food Safety*, 13, 241-260.

493 Nicolai, B. M., Beullens, K., Bobelyn, E., Peirs, A., Saeys, W., Theron, K. I., & Lammertyn, J. (2007).
494 Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review.
495 *Postharvest Biology and Technology*, 46(2), 99-118.

496 Page, D., Van Stratum, E., Degrou, A., & Renard, C. M. G. C. (2012). Kinetics of temperature increase during
497 tomato processing modulate the bioaccessibility of lycopene. *Food Chemistry*, 135(4), 2462-2469.

498 Page, D., Labadie, C., Reling, P., Bott, R., Garcia, C., Gaillard, C., Fourmaux, B., Bernoud-Hubac, N., Goupy,
499 P., George, S., & Caris-Veyrat, C. (2019). Increased diffusivity of lycopene in hot break vs. cold break
500 purees may be due to bioconversion of associated phospholipids rather than differential destruction
501 of fruit tissues or cell structures. *Food Chemistry*, 274, 500-509.

502 Peiris, K. H. S., Dull, G. G., Leffler, R. G., & Kays, S. J. G. G. (1998). Near-infrared (NIR) spectrometric
503 technique for nondestructive determination of soluble solids content in processing tomatoes.
504 *Journal of the American Society for Horticultural Science*, 123, 1089-1093.

505 Peirs, A., Scheerlinck, N., & Nicolaï, B. M. (2003). Temperature compensation for near infrared reflectance
506 measurement of apple fruit soluble solids contents. *Postharvest Biology and Technology*, 30(3), 233-
507 248.

508 Perona, P. (2005). Bostwick degree and rheological properties: an Up-to-date viewpoint. *Applied Rheology*,
509 15, 218–229.

510 Saha, S., Hedau N., Mahajan, V., Singh, G., Gupta, H., & Gahalain, A. (2010). Textural, nutritional and
511 functional attributes in tomato genotypes for breeding better quality varieties. *Journal of the*
512 *Science of Food and Agriculture*, 90(2), 239-244.

513 Santiago, J., Kermani, Z., Xu, F., Van Loey, A.M., Hendrickx M.E. (2017). The effect of high-pressure
514 homogenization and endogenous pectin-related enzymes on tomato purée consistency and serum
515 pectin structure. *Innovative Food Science and Emerging Technologies*, 43, 35–44

516 Scibisz, I., Reich, M., Bureau, S., Gouble, B., Causse, M., Bertrand, D., & Renard, C. M. G. C. (2011). Mid-
517 infrared spectroscopy as a tool for rapid determination of internal quality parameters in tomato.
518 *Food Chemistry*, 125(4), 1390-1397.

519 Shi, H., & Yu, P. (2017). Comparison of grating-based near-infrared (NIR) and Fourier transform mid-
520 infrared (ATR-FT/MIR) spectroscopy based on spectral preprocessing and wavelength selection for
521 the determination of crude protein and moisture content in wheat. *Food Control*, 82, 57-65.

522 Shi, H., Lei, Y. Prates, L. & Yu, P. (2019). Evaluation of near-infrared (NIR) and Fourier transform mid-
523 infrared (ATR-FT/MIR) spectroscopy techniques combined with chemometrics for the determination
524 of crude protein and intestinal protein digestibility of wheat. *Food Chemistry*, 272, 507-513.

525 Sujka, K., Koczon, P., Ceglinska, A., Reder, M. & Ciemniowska-Zytkiewicz, H. (2017). The application of FT-
526 IR spectroscopy for quality control of flours obtained from polish producers. *Journal of Analytical*
527 *Methods in Chemistry*, 1-9.

528 Svelander, C. A., Tiback, E. A., Ahrne, L. M., Langton, M., Svanberg, U.S.O., & Alming, M.A.G. (2010).
529 Processing of tomato: impact on in vitro bioaccessibility of lycopene and textural properties. *Journal*
530 *of the Science of Food and Agriculture*, 90(10), 1665-1672.

531 Talari, A. C. S., Martinez, M. A. G., Movasaghi, Z., Rehman, S., & Rehman, I. U. (2017). Advances in Fourier
532 transform infrared (FTIR) spectroscopy of biological tissues. *Applied Spectroscopy Reviews*, 52(5),
533 456-506.

534 Thomas, E. V., & Ge, N. X. (2000). Development of robust multivariate calibration models. *Technometrics*,
535 42(2), 168-177.

536 Wilkerson, E. D., Anthon, G. E., Barrett, D. M., Sayajon, G. F. G., Santos, A., & Rodriguez-Saona, L. (2013).
537 Rapid Assessment of Quality Parameters in Processing Tomatoes Using Hand-Held and Benchtop
538 Infrared Spectrometers and Multivariate Analysis. *Journal of Agricultural and Food Chemistry*, 61(9),
539 2088-2095.

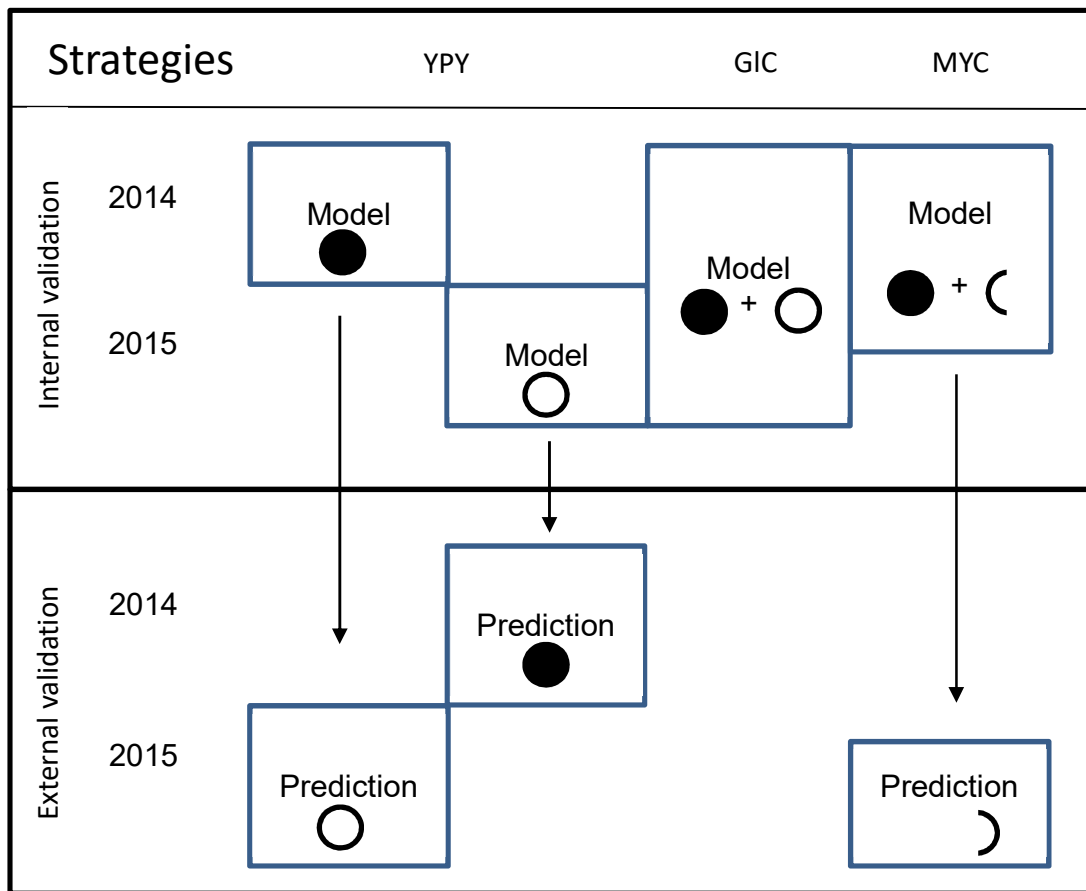
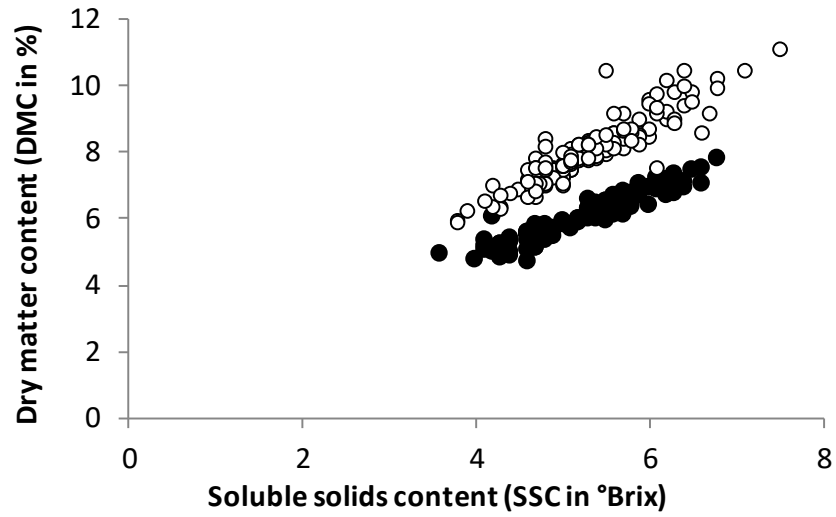


Figure 1. The three tested strategies to build models using both mid-infrared spectra and reference data of quality traits.

with YPY; Year per Year models, GIC: Global Combining models and MYC: Multi-Year Combining models.

A)



B)

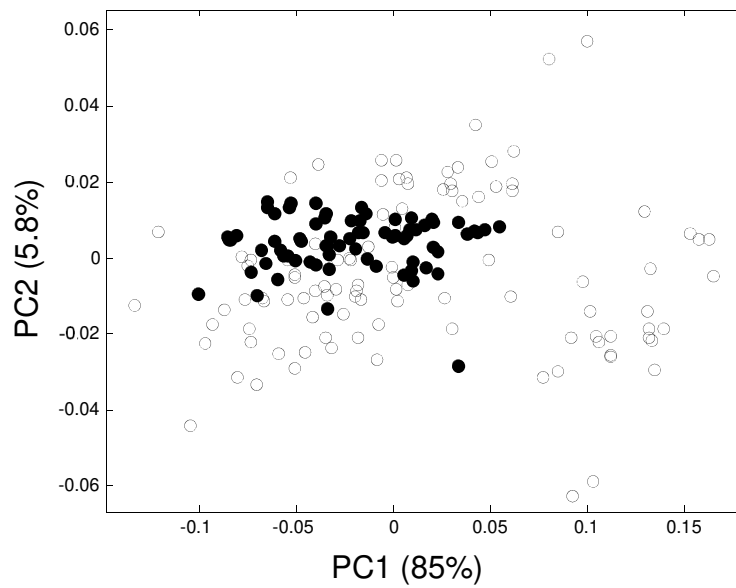


Figure 2. Variability of processing type tomatoes in France over two years. A) Biplot between soluble solids content (SSC) and dry matter content (DMC) and (B) Principal Component Analysis (PCA) performed on spectral data ($2000-900\text{ cm}^{-1}$) with tomatoes characterized in 2014 (●) and in 2015 (○).

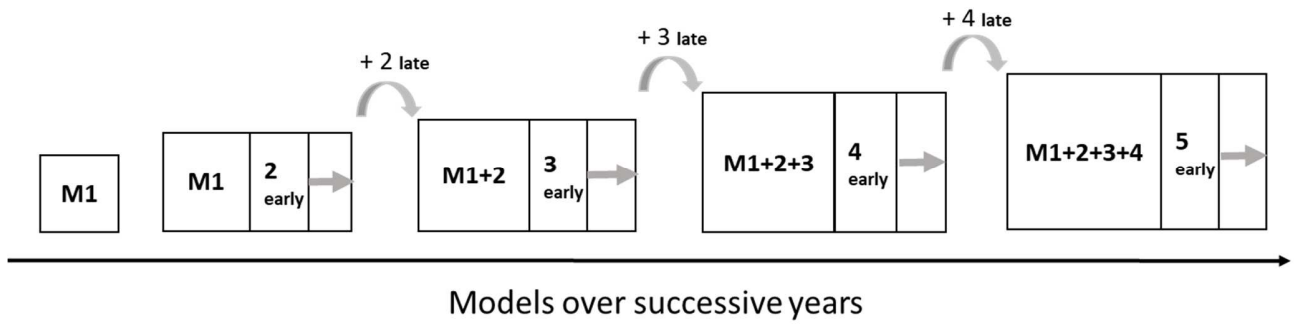
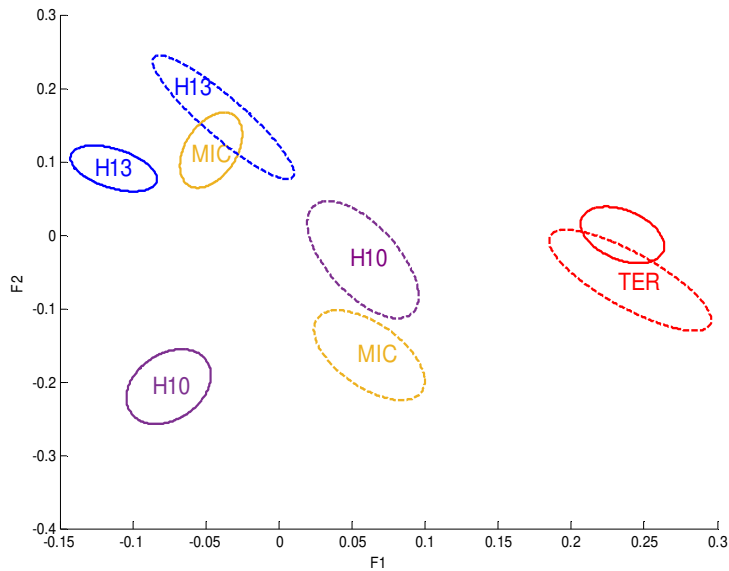


Figure 3. The best strategy to improve model ability over successive years.

With 2, 3, 4, 5: early and late data each year added in the previous models identified by M1+2+... and arrows simulating the model use to predict firstly the late tomato quality traits each year and then the tomatoes of next years.

A.



B.

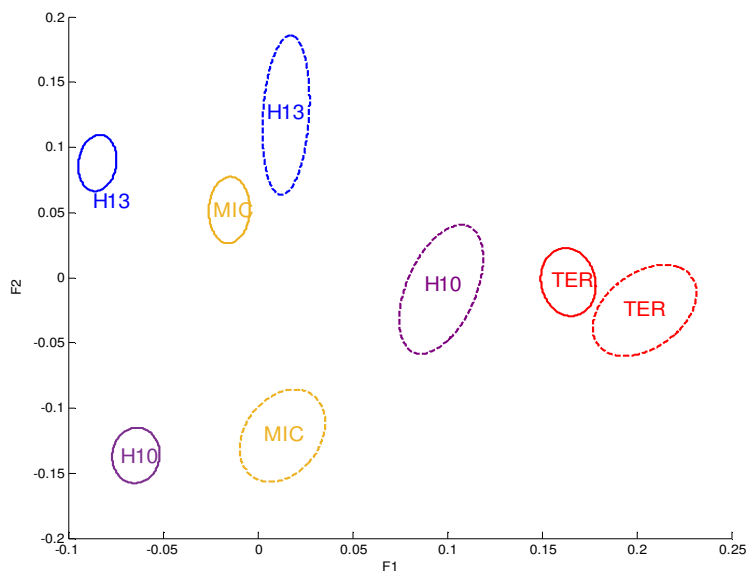


Figure 4. Factorial Discriminant Analysis (FDA) maps performed on mid-infrared spectral (MIRS) data ($1200-900\text{ cm}^{-1}$) of fresh tomato (A) and processed purees (B). Ellipses drawn with a P value of the confidence interval of 0.05, with continuous line (2016) and dotted line (2017). Terr: Terradou, H10: H1015, H13: H1311 and MIC: Miceno. FDA was calculated on 2016 data. 2017 data were added as illustrated data.

Table 1. Soluble solids content (SSC), dry matter content (DMC) and titratable acidity (TA) of fresh tomatoes measured by reference methods over two successive years.

Quality traits	Year	Mean	SD	Min	Max
SSC (°Brix)	2014	5.2	0.7	3.6	6.8
	2015	5.4	0.7	3.8	7.5
DMC (%)	2014	6.0	0.7	4.7	7.8
	2015	7.9	1.1	5.7	11.1
TA (mmol H ⁺ /Kg)	2014	59.1	6.7	45.3	76.8
	2015	55.6	10.6	30.2	81.7

SD: standard deviation, n=102 samples in 2014 and 144 in 2015 and each sample was a homogenate of 15 tomato fruits.

Table 2. Validation results to compare the performance of the models to predict SSC, DMC and TA depending on the strategies.

Quality trait	Sampling	LV	Calibration		Validation		RPD
			R_c^2	RMSE _c	R_v^2	RMSE _v	
SSC (°Brix)	YPY (2014)	7	0.95	0.16	0.95	0.16	4.3
	YPY (2015)	4	0.96	0.14	0.94	0.13	3.6
	GIC	8	0.95	0.15	0.93	0.17	3.9
	MYC	9	0.97	0.13	0.95	0.14	4.5
DMC (%)	YPY (2014)	7	0.89	0.24	0.87	0.25	2.6
	YPY (2015)	4	0.82	0.39	0.85	0.41	2.6
	GIC	6	0.86	0.48	0.85	0.49	2.5
	MYC	10	0.92	0.37	0.94	0.34	3.9
TA (mmol H ⁺ /Kg)	YPY (2014)	10	0.89	2.33	0.84	2.19	2.5
	YPY (2015)	10	0.97	2.01	0.96	2.23	4.6
	GIC	9	0.84	3.74	0.90	3.47	3.1
	MYC	9	0.84	3.92	0.79	4.15	2.2

LV: latent variables, R^2 : coefficient of determination, RMSE: root mean square error, with _c for calibration and _v for validation; RPD: ratio of the standard deviation (SD) of the response variable in the validation set to the RMSE_v; Strategies named: YPY for year-per-year models, GIC : global combining models, MYC : Multi-year combining models

Sample number was n=102 in 2014, n=144 in 2015, n=246 in 2014 + 2015 and n=181 in 2014 + 2015 early (all samples in 2014 and until August, 18th 2015).

With 2014 and 2015 from the Scenario1 (YPY models); 2014+2015 from the Scenario 2 (GIC) and 2014 + early 2015 from the Scenario 3 (MYC models combining all data of the first year 2014 and data of the beginning of the second year until August, 18th 2015).

Table 3. External validation results to compare the performance of the models to predict SSC, DMC and TA depending on the strategies

Quality trait	Models	Predicted samples	External validation		
			R_p^2	RMSE _p	RPD
SSC (°Brix)	YPY (2014)	2015	0.31	0.90	0.5
	YPY (2015)	2014	0.88	0.25	2.7
	MYC	end 2015	0.95	0.11	4.3
DMC (%)	YPY (2014)	2015	0.13	3.89	0.3
	YPY (2015)	2014	0.79	2.54	0.1
	MYC	end 2015	0.81	0.36	2.8
TA (mmol H ⁺ /Kg)	YPY (2014)	2015	0.27	33.12	0.3
	YPY (2015)	2014	0.16	53.44	0.3
	MYC	end 2015	0.88	2.81	2.1

R_p^2 : coefficient of determination of external validation, RMSE_p: root mean square error of external validation.

With 2014 and 2015 from the Scenario1 (YPY models) with n=102 in 2014 and n=144 in 2015; 2014 + early 2015 from the Scenario 3 (MYC models) with n=181 in 2014 + 2015 early (all samples in 2014 and data of the beginning of the second year until August, 18th 2015) and n=65 in end 2015 (data from August, 18th 2015).

Table 4. Matrices of confusion given by the Factorial discriminant analysis (FDA) using PC scores of the PCA (Principal Component Analysis) performed on the spectral data (2000-900 cm⁻¹) of the fresh tomato homogenates and their corresponding cooked purees. Three factors were tested with A: years, B: type of samples and C: varieties.

A. Year

	2016	2017
2016	143	1
2017	0	192

B. Type of tomato-based products

	CB	FR	HB
CB	94	1	1
FR	0	144	0
HB	6	1	89

C. Variety

	H10	H13	MIC	TER
H10	45	9	24	6
H13	12	59	13	0
MIC	19	9	56	0
TER	3	1	4	76

The total number of samples for each condition being 2016: 144 samples; 2017: 192 samples; CB: 96 samples; HB: 96 samples and fresh: 144 samples; 84 samples for each of the H10, H13, MIC and TER varieties.

Table 5. Prediction of quality traits of industrial products using both reference data acquired by laboratory measurements and by plant control quality.

Samples	Quality traits	Reference data		LV	Calibration		Cross-validation		
		Mean	SD		R_c^2	$RMSE_c$	R_{cv}^2	$RMSE_{cv}$	RPD
All samples	SSC (°Brix)	17.4	8.4	5	0.99	0.73	1.00	0.66	12.7
	TA (mmol H ⁺ /Kg)	131.2	71.6	7	0.99	8.79	0.98	9.47	7.6
	DMC (%)	20.5	9.4	5	0.99	1.01	0.99	0.99	9.4
Juices and purees	SSC (°Brix)	11.4	2.2	5	0.92	0.67	0.87	0.68	3.2
	TA (mmol H ⁺ /Kg)	93.9	22.4	8	0.88	8.95	0.82	8.82	2.5
	DMC (%)	15.5	3.1	4	0.90	1.10	0.76	1.13	2.7
All samples	SSC (°Brix)	15.0	7.2	5	0.99	0.98	0.99	0.97	7.5
	pH	4.4	0.1	9	0.55	0.09	0.51	0.10	1.4
	Bw	6.0	5.0	5	0.80	1.02	0.77	1.08	4.6
Juices and purees	SSC (°Brix)	11.5	2.5	7	0.92	0.67	0.93	0.70	3.6
	pH	4.4	0.2	5	0.59	0.10	0.38	0.13	1.3
	Bw	2.8	1.5	6	0.46	1.07	0.35	1.22	1.2

When all samples (juices, purees and pastes) were used: n=76 in the calibration set and n=38 in the cross-validation set. When only juices and purees were used: n=57 in the calibration set and n=28 in the cross-validation set.

LV: latent variables, R²: coefficient of determination, RMSE: root mean square error, with _c for calibration and _v for validation; RPD: ratio of the standard deviation (SD) of the response variable in the validation set to the RMSE_v. Bw: Bostwick.