ARTIFICIAL INTELLIGENCE ATLAS SESSION

EXPLAINABLE MACHINE LEARNING PREDICTION OF PCB-138 BEHAVIOR PATTERNS IN EDIBLE FISH FROM CROATIAN ADRIATIC

Andreja Stojić[1,2]*,

Bosiljka Mustać[3],

Gordana Jovanović[1,2]

[1]Institute of Physics Belgrade, National
 Institute of the Republic of Serbia,
 University of Belgrade,
 Belgrade, Serbia

[2]Environment and Sustainable Development,
 Singidunum University,
 Belgrade, Serbia

[3]Department of ecology, agronomy and
 aquaculture,
 University of Zadar,
 Zadar, Croatia

Abstract:

Fish consumption, especially consumption of oily marine species, is globally increasing since it has been recommended by dieticians due to high content of polyunsaturated ω-3 and ω-6 (PUFAs) fatty acids in fish tissue. Health benefits of PUFA ingestion coincide with the risk of intake of hazardous lipophilic persistent pollutants including organochlorine pesticides (OCPs) and related polychlorinated biphenyls (PCBs). We examined the impacts of 18 fatty acids (FAs) and 36 toxic organic and inorganic contaminants on the behavior patterns of indicator congener PCB-138 in marine fish using eXtreme Gradient Boosting (XGBoost), SHapley Additive exPlanations (SHAP), and SHAP value fuzzy clustering. XGBoost indicated non-linear relationships between investigated variables that surpasses indications suggested by commonly applied correlation matrices. Ten extracted fuzzy clusters of SHAP values revealed that higher intake of saturated myristic-C14:0 and margaric-C17:0 acids followed by intake of nutritionally beneficial eicosadienoic acid (C20:2n-6) mostly contributed to the PCB-138 bioaccumulation. Important impacts on PCB-138 behavior patterns were also registered for chemically allied indicator congeners (-153 and -180) and organochlorines' metabolite *p,p'*-DDE. Less prominent were the associations between target congener and the most toxic dioxin-like PCBs.

1. INTRODUCTION

Anthropogenic activities have led to several global, regional and local environmental issues related to air, water and soil pollution, gradual decrease of the stratospheric ozone layer, decrease in biodiversity, etc. Organic and inorganic contaminant emissions and dispersion caused by human activities refer to different classes of polychlorinated biphenyls (PCBs), polycyclic aromatic pollutants, trace metalsand natural radio-activity. Organochlorine pesticides (OCPs) and PCBs are well-known as persistent organic pollutants (POPs). They are long-lived contaminants and possess numerous adverse effects on living organisms including humans and animals. Because of that their production and use is limited

Correspondence:

Andreja Stojić

e-mail:
andreja.stojic@ipb.ac.rs

23

or forbidden in most countries. POPs bind tightly to particles in soil and sediment, which can act as a secondary sources of contamination for environmental media (water, air and living organisms).

At a global level, the oceans are the final sink/destination of POPs. The oceans act as a secondary source of contamination because POPs are slowly degraded and bioaccumulated in marine organisms, which are at the bottom of the food chain and represent a source of chemical hazards in human nutrition. The assessment of environmental exposure to marine toxins is based on POPs concentration data in the samples of water, plants and food. Small pelagic oily fish are highly recommended nutrient source worldwide due to their content of protein, minerals and healthy fats including omega-3 (ω-3) and 6 ω-6) polyunsaturated fatty acids (PUFAs) [1]. Benefits of the fatty acid consumption are associated with normal growth and development, the prevention of cardiovascular and inflammatory diseases, as well as cognitive decline and dementia. However, these FAs represent a very suitable matrix for bioaccumulation of highly lipophilic xenobiotics such as POPs. Although FAs profile and organochlorines content have been evaluated in numerous marine fish species worldwide, there are few data on their interrelations.

In this study, the presence of OCPs and PCBs has been investigated in small pelagic edible fish species: sardine *Sardina pilchardus* (Walbaum, 1792), anchovy *Engraulis encrasicolus* (Linnaeus, 1758), round sardinella *Sardinella aurita* (Valenciennes, 1847), chub mackerel *Scomber japonicus* (Houttuyn, 1782) and horse mackerel *Trachurus trachurus* (Linnaeus, 1758). We applied eXtreme Gradient Boosting (XGBoost), SHapley Additive exPlanations (SHAP), and SHAP value fuzzy clustering aiming to obtain a detailed insight into the distribution of indicator congener PCB-138 in the fish species. The impacts of the following factors including the level of OCPs, PCBs, saturated fatty acids (SFAs), monounsaturated fatty acids (MUFAs), PUFAs and heavy metals were evaluated by SHAP since the method offers uniquely consistent and locally accurate solutions that have been confirmed in the previous investigations of environmental phenomena [2].

2. MATERIAL AND METHODS

Sampling

Fish samples were collected along the eastern Croatian Adriatic Sea during 2014 and 2016. Details about sampling were previously described [3]. Total of 107 fish samples were collected in various fisheries coastal (A, E, and F) and off-coast (B and C) zones. Approximately 50 specimens was sampled randomly using purse seine catches (mesh size: 8 mm/bar length) totaling 107 pooled samples from a fillet of specimen.

Chemical analyses
POPs

Chemical analysis of POPs was previously described elsewhere [3]. In brief, seven OCPs (HCBα-, β-, and γ-HCH, *p,p'*-DDT, *p,p'*-DDE, and *p,p'*-DDD), six indicator PCB congeners (PCB-28, PCB-52, PCB-101, PCB-138, PCB-153, and PCB-180), eight mono ortho congeners (PCB-105, PCB-114, PCB-118, PCB-123, PCB-156, PCB-157, PCB-167, PCB-189) and () PCB-60, PCB-74 and PCB-170] were analysed. High-resolution gas chromatography with electron capture detector (s) was applied for the compound identification and details are described previously [3].

The recoveries for the PCBs was in the range between 75% to 89% while relative standard deviation (RSD) between 1% to 11% was obtained. The recoveries for OCPs were in the range from 76% to 86%, with RSD from 1% to 11%. for both PCBs and OCPs, the determination limits were 0.01 ng g−1 of fresh weight.

Elements

Sixteen macro- and micro-elements including toxic metals (Na, Mg, K, Ca, As, Cd, Co, Cr, Cu, Mn, Fe, Hg, Ni, Zn, Pb and Se) were analyzed. Homogenized fish samples (0.5 g) were basted by 5 mL of nitric acid (67% TraceMetal grade, Fisher Scientific, Bishop, UK) and 1.5 mL of hydrogen peroxide (30% analytical grade, Sigma-Aldrich, St. Louis, MA, USA). Afterwards, microwave digestion was tperformed as previously described [4].Analysis of the elements was performed by inductively coupled plasma mass spectrometry (ICP-MS), (iCap Q mass spectrometer, Thermo Scientific, Bremen, Germany). The most abundant isotopes were used for quantification.

For five-point calibration, solutions of Fe, Zn, Cu, Mn, Se, Cr, Co, Ni, Na, K, Mg and Ca were prepared in the concentration range of 0.2–2.0 mg L^{-1}. The concentration of calibration-solution forCd, Hg and As was in the range of between 0.2 and 2.0 µg L^{-1} and for Pb in the range between 2.0 and 20.0 µg L^{-1}.

To check for the accuracy of the analysis, the certified reference material NIST SRM 1577c (bovine liver, Gaithersburg, MD, USA) was analyzed in the same manner as the fish samples. For all elements, the obtained results were within the satisfactory range of the certified values

Fatty acids

The concentrations of 6 SFAs (myristic, pentadecylic, palmitic, palmitoleic, margaric, and stearic acid), 3 MUFAs (oleic, paullinic and arachidonic acid) and 9 PUFAs, ω-6 and ω-3 families (linoleic, α-linolenic, icosadienoic, dihomo-γ-linolenic, eicosatrienoic, arachidonic, eicosapentaenoic, docosapentaenoic and docosahexaenoic) were determined according to procedure given in detail previously [5]. Prior analysis, the samples were partially thawed at +4 °C. Accelerated solvent extraction (ASE 200, Dionex, Sunnyvale, CA) using a mixture of n-hexane and iso-propanol (60:40 v/v) were applied for determination of total lipids for fatty acid content The samples were analysed as FAME (fatty acis methyl esters) by gas–liquid chromatography (Shimdzu, Japan). Flame ionisation detector (GC/FID) and fused silica cianopropyl HP-88 column were used. The chromatographic peaks were identified and quantified using Supelco 37 Component mix standard and internal standard (heneicosanoic acid methyl ester), respectively.

Data analysis

The relationships between PCB-138 and all other measured parameters were modeled by the XGBoost regression. The details on the method are given elsewhere [6]. In this study, we used Python XGBoost implementation. The dataset was split into training (80%) and validation (20%) sets. Hyperparameter tuning was implemented using a brute-force grid search and 10-fold stratified cross-validation. The best performing hyperparameter values were used for the final model.

The explainability of the produced XGBoost model that operates with high-dimensional input data in a non-linear fashion was obtained by using explainable artificial intelligence method SHapley Additive exPlanations (SHAP) [7]. Based on the game theory, the Shapley explanations represent the only possible locally accurate and globally consistent feature attribution values. In this study, we applied the fuzzy clustering of absolute SHAP attributions to identify and characterize the relations among the measured parameters responsible for PCB-138 behavior.

2. RESULTS AND DISCUSSION

Pollutant toxicological profile

Small edible pelagic fishes live short at the bottom of the marine food chain (plankton<sardine species, anchovy<mackerel species), which are expected to bioaccumulate low levels of environmental contaminants. The transfer of accumulated hazards to the human body via fish intake is generally consider to be low. As shown by study results, inorganic compounds, macro-elements and toxic heavy metals, were prominently more abundant in the fish tissue than organic xenobiotics, OCPs and PCBs.

Biological effects of inorganic elements depend on the processes including absorption, accumulation, elimination and biotransformation into less or more toxic metabolites. Several minutes from absorption in the gut, the elements reach internal organs such as heart, liver, kidney and brain, while their penetration to muscles and adipose tissue occurs more slowly, up to several hours [8]. However, when they are present in low concentrations in the aquatic surrounding, their accumulation is less frequent as well. In the studied fish species, the concentrations of macro- (K > Na > Ca > Mg) and micro-elements (Fe > Zn > Cu) were within the normal physiological range, below the levels that could potentially cause pathological changes in tissues and organs, because of the absence of dominant well-known sources of pollution. In addition, the levels of the following elements: Hg, Pb and Cd, were lower than the maximum thresholds prescribed by the existing EU regulation [9].

Among POPs, p,p'-DDE, PCB-153, PCB-138 and PCB-180 were the most dominant. Although it is to be expected that larger, long-lived species, such as chub mackerel and horse mackerel, that are at higher trophic level, will uptake more organic contaminants than sardine species and/or anchovy, the values of the toxicological parameters such as the total quantity of indicator PCBs and WHO-dioxin-like PCBs toxic equivalents were below threshold concentrations of 75 ng g^{-1} w.w. and 6.5 pg g^{-1} [10]. As can be concluded, the examined fish speies appeared to be safe for human diet with regards to the presence of toxic chemicals.

PCB-138 patterns

As represented by Pearson's correlation analysis, significant linear correlation coefficients (r>0.90) were found between the following pairs of the investigated variables: p,p'-DDE–PCB-118; p,p'-DDE–PCB-138; p,p'-DDD–PCB-105; p,p'-DDD–PCB-180; PCB-138–

PCB-118; PCB-138–PCB-153; PCB-153–PCB-118; PCB-153–PCB-170 and PCB-156–PCB-180. The correlations between the POPs pairs: γ-HCH–PCB-170; γ-HCH–PCB-156; γ-HCH–PCB-118; γ-HCH–PCB-114; γ-HCH–PCB-180; *p,p*'-DDE–PCB-153; *p,p*'-DDT–PCB-170; *p,p*'-DDT–PCB-156; *p,p*'-DDT–PCB-118; *p,p*'-DDT–PCB-153; PCB-138–PCB-170; PCB-180–PCB-170; PCB-180–PCB-156; PCB-105–PCB-180; PCB-105–PCB-156 and PCB-105–PCB-170 were in the range from 0.80 to 0.90. As can be seen, the listed species with similar chemical structure and common origin display similar behavior patterns, as discussed below. We assumed that other methods apart from commonly applied correlation matrices could be further employed to describe the associations between POPs and FAs or inorganic contaminants in more details.

In this study, XGBoost has been successfully employed for the investigation of non-linear relationships between PCB-138 levels and key variables that shape its' behavior pattern in marine fish tissue. The predicted/observed calculated relative error was 11.2%, while the r^2 were 0.99 (*Fig. 1*).

Fuzzy clustering of SHAP values extracted ten groups (relative error < 10%) of similar variables that shape the dynamics of PCB-138 (*Fig. 2*. As labelled by red color, the constituents of six clusters, C2 and C6–C10, strongly positively impacted PCB-138 concentrations higher than 0.5 ng g^{-1} and up to 1.5 ng g^{-1}. Dominant influences are attributed to the higher levels of myristic acid

(> 7%) and the following compounds: *p,p*'-DDE, PCB-101, PCB-118, PCB-123, PCB-153, and PCB-180, as well as to lower margaric content (≈ 1%). Although indicator congeners (-28, -52, -101, -138, -153, and -180), which are often classified as non-dioxin-like PCBs, dominated over dioxin-like PCBs, both classes were of significance when evaluating PCB-138 patterns. The relationships between PCB congeners and their presence at higher concentration levels are due to their structure referring to molecule rigidity and the number of attached halogen atoms, and consequently, perseverance and prolonged half-lives in marine environment.

Out of 18 investigated FAs, saturated myristic (C14:0) and margaric (C17:0) acid mostly contributed to the PCB-138 bioaccumulation followed by nutritionally beneficial eicosadienoic (C20:2n-6) and dihomo-γ-linolenic acid (C20:2n-3).
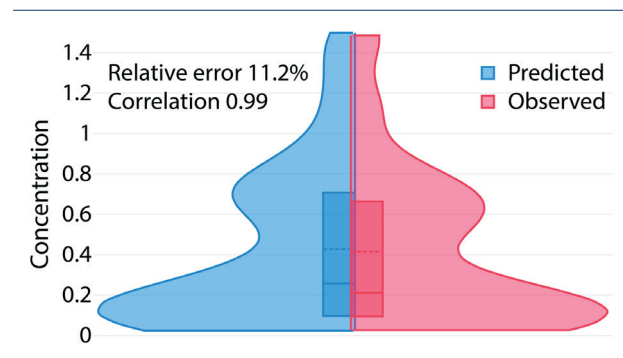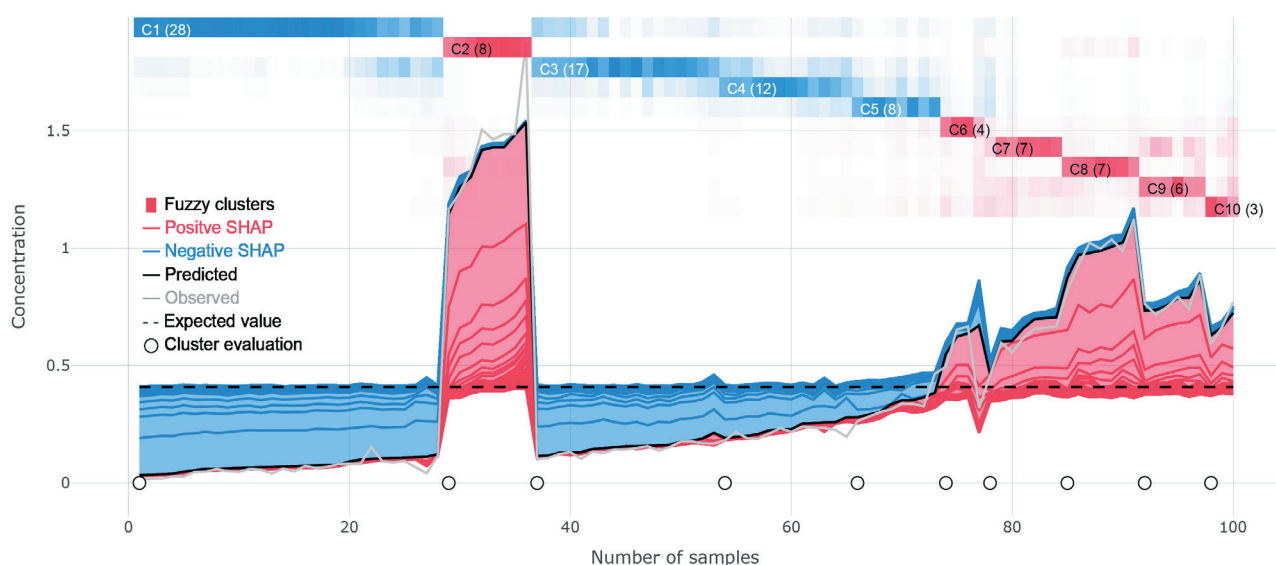

Fig. 1. XGBoost evaluation statistics.

Fig. 2. PCB-138 SHAP force plot.

We have noted that these fatty acids were not the most abundant in the studied fish species, while the dominance of other FAs was recorded in the following order: palmitic > oleic > docosahexaenoic > stearic > myristic > eicosapentaenoic > palmitoleic > linoleic acid. Saturated acids such as C14:0 and C17:0 dominantly occured as esterified polar phospholipids and therefore, SFAs and associated contaminants are more bioavailable than FAs presented in non-esterified free form. Eicosapentaenoic acid (EPA, 20:5n-3) and docosahexaenoic acid (DHA, 22:6n-3), are widely known as the most nutritionally relevant ω-3 PUFAs in oily blue fish, such as mackerel, sardines and anchovies. However, they appear to have no impact on the PCB-138 accumulation. Chub mackerel has been known to contain high percentage of the PUFAs in the form of free FAs, which are less efficiently absorbed than other lipid classes and easily eliminated from human intestines where they are dispersed into mixed micelles and bound to soluble lipid-binding proteins [11]. In addition, PUFAs are more susceptible to oxidative degradation when found as free FAs.

Four clusters (C1 and C3–C5, ≈70%) dominantly showed negative correlations with PCB-138 concentrations below 0.5 ng g^{-1}. The same parameters as discussed above shaped the clusters and negative relationships imply that low concentrations of SFAs do not add to PCB-138 uptake, or uptake of POPs from different sources, that are found to have variable impact, depending on the sampling time, seasonal and other fishery zone-related factors.

The plot in *Fig. 3.* shows the impact which each cluster has on output of the model. High levels of the cluster (C2, C6-C10) constituents, primarily PCB-153, *p,p*'-DDE, PCB-118, PCB-170, PCB-180 and myristic acid, has a high and positive impact on the PCB-138 patterns, as shown by right-oriented long distribution tail. Lower concentrations of these variables (< 0.1 ng g^{-1}) are negatively correlated with the target variable, as indicated by negative relative SHAP value.
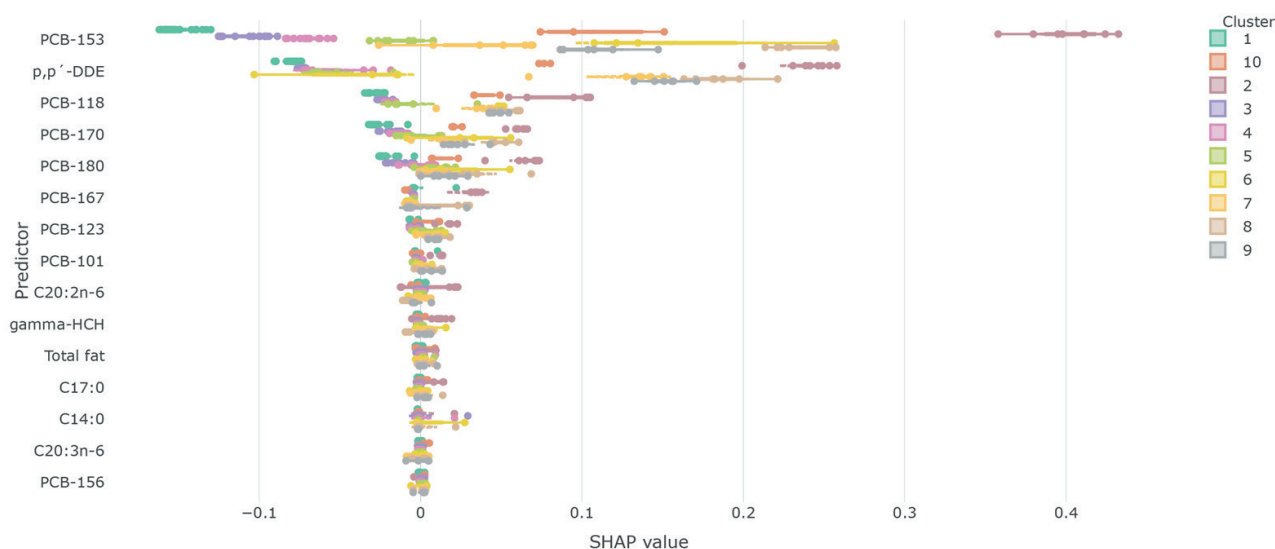

Fig. 3. PCB-138 SHAP value cluster evaluation.

4. CONCLUSION

Small edible fish species could be considered as both nutritionally valuable food source and sources of hazardous organochlorine pollutants, which negatively affect human health. Significant indications on relationships between POPs uptake and fatty acid in fish tissue has been reported worldwide. In this study, we presented a promising methodology, an explainable artificial

intelligence methods (XGBoost and SHAP), which was used with the aim of gaining better understanding of specific interrelations between fatty acid content and contaminants in consumable marine fish species. Out of 18 fatty acids, two saturated (myristic and margaric) and two ω-3 and 6 (eicosadienoic and dihomo-γ-linolenic) acids were identified as crucial for the bioaccumulation of PCB-138 in sardine, anchovy and mackerel species. However, nutritionally beneficial EPA and DHA are

assumed to have no impacts on the contaminant uptake. The content of macro-elements and heavy metals is not related to PCB-138 chemodynamics in fish tissue, while influences of *p,p'*-DDE and both indicator and toxicologically congeners (-101, -118, -123, -153, and -180) were evident. Finally, methods have been successfully verified as a reliable means for examination of the relationships between POPs and FAs that overpowers commonly employed statistical approaches.

ACKNOWLEDGMENT

REFERENCES

[1] Food and Agriculture Organization of the United Nations (FAO), "The state of world fisheries and aquaculture worldwide 2018 - Meeting the sustainable development goals" 2018, pp. 1-227, ISBN 978-92-5-130562-1.

[2] A. Stojić, N. Stanić, G. Vuković, S. Stanišić, M. Perišić, A. Šoštarić, and L. Lazić, "Explainable extreme gradient boosting tree-based prediction of toluene, ethylbenzene and xylene wet deposition." Sci. Total Environ., 2019, pp. 140-147.

[3] G. Vuković, S. Herceg Romanić, Ž. Babić, B. Mustać, M. Štrbac, I. Deljanin, D. Antanasijević, "Persistent organic pollutants (POPs) in edible fish species from different fishing zones of Croatian Adriatic." Marine Pollut. Bull. 2018, pp. 71-80.

[4] J.M. Đinović-Stojanović, D.M. Nikolić, D.V. Vranić, J.A. Babić, M.P. Milijašević, L. Pezo, S. Janković "Zinc and magnesium in different types of meat and meat products from the Serbian market." J. Food Compost. Anal., 2017, pp. 50-54.

[5] A. Špirić, D. Trbović, D. Vranić, J. Đinović, R. Petronijević, V. Matekalo-Sverak, "Statistical evaluation of fatty acid profile and cholesterol content in fish (common carp) lipids obtained by different sample preparation procedures.", Anal. Chim. Acta, 2010, pp. 66-71.

[6] T. Chen, C. Guestrin, "Xgboost: A scalable tree boosting system", In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016, pp. 785-794.

[7] S. Lundberg, S. Lee, "A unified approach to interpreting model predictions". In Adv. Neural Inf. Process. Syst. 2017, pp. 4765-4774.

[8] E. Brucka-Jastrzêbska, D. Kawczuga, M. Rajkowska, M. Protasowick, "Levels of microelements (Cu, Zn, Fe) and macroelements (Mg, Ca) in freshwater fish." J. Elementol., 2009, pp. 437-447.

[9] European Commission, 2006. Commission Regulation (EC) No 1881/2006 of 19 December 2006 setting maximum levels for certain contaminants in foodstuffs 20.12.2006 Off. J. Eur. Union, pp. L 364, 5-24.

[10] European Commission, 2011. Commission Regulation (EU) No 1259/2011 of 2 December 2011 amending Regulation (EC) No 1881/2006 as regards maximum levels for dioxins, dioxin-like PCBs and non-dioxin-like PCBs in foodstuffs. Off. J. Eur. Union, pp. L 320, 18-23.

[11] M. Á. Rincón-Cervera, V. González-Barriga, R. R. Valenzuela, S. López-Arana, J. Romero, A. Valenzuela, "Profile and distribution of fatty acids in edible parts of commonly consumed marine fishes in Chile." Food Chem, 2019, pp. 123-129.