Traballo Fin de Máster

Estudo e aplicación de modelos estatísticos para a estimación da variabilidade do Carbono Orgánico Disolto no océano global

Mateo Soto Tielas

Máster en Técnicas Estatísticas

Curso 2024-2025

Proposta do Traballo Fin de Máster

Título en galego: Estudo e aplicación de modelos estatísticos para a estimación da variabilidade do Carbono Orgánico Disolto no océano global
Título en español: Estudio y aplicación de modelos estadísticos para la estimación de la variabilidad del Carbono Orgánico Disuelto en el océano global
English title: Study and application of statistical models for estimating the variability of Dissolved Organic Carbon in the global ocean
Modalidad: Modalidad B
Autor/a: Mateo Soto Tielas, Universidade de Vigo
Director/a: Marta Sestelo Pérez, Universidade de Vigo; Nora Martínez Villanueva, Universidade de Vigo
Tutor/a: Xosé Antonio Padín Álvarez, Instituto de Investigacións Mariñas (IIM-CSIC); Marcos Morente Fontela, Instituto de Investigacións Mariñas (IIM-CSIC)
Breve resumo do traballo: Este traballo de fin de mestrado ten como obxectivo principal a creación e avaliación de modelos estatísticos para a estimación e xeración de climatoloxías do Carbono Orgánico Disolto no océano. O estudo levarase a cabo a escala global co fin de identificar patróns e comportamentos relevantes que contribúan a unha mellor comprensión da dinámica do Carbono Orgánico Disolto en distintos contextos xeográficos do oceáno global.
Recomendacións:
Outras observacións:

Dona Marta Sestelo Pérez, Profesora Titular da Universidade de Vigo, dona Nora Martínez Villanueva, Profesora Axudante Doutora da Universidade de Vigo, don Xosé Antonio Padín Álvarez, Investigador principal de Instituto de Investigacións Mariñas (IIM-CSIC), e don Marcos Morente Fontela, Persoal técnico de apoio de Instituto de Investigacións Mariñas (IIM-CSIC), informan que o Traballo Fin de Máster titulado

Estudo e aplicación de modelos estatísticos para a estimación da variabilidade do Carbono Orgánico Disolto no océano global

foi realizado baixo a súa dirección por don Mateo Soto Tielas para o Máster en Técnicas Estatísticas. Estimando que o traballo estea terminado, dan a súa conformidade para a presentación e defensa ante un tribunal.

En Ponteareas, a 03 de Xuño de 2025.

A directora:
Dona Marta Sestelo Pérez

SESTELO PEREZ MARTA - 76998300G
Firmado digitalmente por SESTELO PEREZ MARTA - 76998300G
Fecha: 2025.06.03 18:22:25 +02'00'

A directora:
Dona Nora Martínez Villanueva

MARTINEZ VILLANUEVA NORA - 53179846M
Firmado digitalmente por MARTINEZ VILLANUEVA NORA - 53179846M
Fecha: 2025.06.03 18:24:18 +02'00'

O tutor:
Don Xosé Antonio Padín Álvarez

PADIN ALVAREZ JOSE ANTONIO - DNI 76864913V
Digitally signed by PADIN ALVAREZ JOSE ANTONIO - DNI 76864913V
Date: 2025.06.03 18:31:47 +02'00'

O tutor:
Don Marcos Morente Fontela

MORENTE FONTELA MARCOS - DNI 71655386J
Firmado digitalmente por MORENTE FONTELA MARCOS - DNI 71655386J
Fecha: 2025.06.03 20:15:36 +02'00'

O autor:
Don Mateo Soto Tielas

Firmado por SOTO TIELAS MATEO - ***7631** el día 03/06/2025 con un certificado emitido por AC FNMT Usuarios

---

Declaración responsable. Para dar cumplimento á Lei 3/2022, do 24 de febreiro, de convivencia universitaria, referente ao plaxio no Traballo Fin de Máster (Artigo 11, Disposición 2978 do BOE núm. 48 de 2022), **o autor declara** que o Traballo Fin de Máster presentado é un documento orixinal no que se tiveron en conta as seguintes consideracións relativas ao uso do material de apoio desenvolto por outros/as autores/as:

- Todas as fontes empregadas na elaboración deste traballo foron citadas correctamente (libros, artigos, apuntes de profesorado, páxinas web, programas,...)
- Calquer contido copiado ou traducido textualmente púxose entre aspas, citando a súa procedencia.
- Fíxose constar explicitamente cando un capítulo, sección, demostración,...sea unha adaptación casi literal dalgunha fonte existente.

E, acéptase que, se se demostrara o contrario, se lle aplicarán as medidas disciplinarias que correspondan.

Agradecementos

Primeiro, agradecer a Toni e Marcos do IIM-CSIC por abrirme as portas do centro e do seu laboratorio dende o primeiro día. Grazas por ofertarme esta temática e por acompañarme durante todo este proceso.

A continuación, ás fantásticas titoras Marta e Nora: polas numerosas reunións que tivemos, o tempo que investiron en min e por preocuparse moito máis aló do traballo académico. Grazas, sodes unhas referentes e merecedes todo o bo que vos pase cara o futuro.

Tamén a toda a xente do meu carón. Somos quen somos pola xente que nos rodea e non podo estar máis agradecido. Grazas por elixir camiñar xuntos nesta vida; aínda que non cambiemos este mundo cruel, facémolo moito máis agradable.

Á miña compañeira, Irene. Grazas por ser o meu fogar, por ensinarme tanto e por acompañarme do xeito no que o fas. Nunca imaxinei ter tanta sorte. Despois deste traballo deixarei de "xogar tanto coa maquinita", unha mágoa.

Case para o final, ao resto da miña familia por axudarme en todo este camiño e por converterme na persoa que son agora. Ma e Pa, Dani e á avoa, grazas e unha forte aperta.

Por último, agradecer a todas as persoas, profesores, adestradores, compañeiros... que me axudaron a construírme como persoa, a formarme e a ter paixón pola vida, pola cultura e pola lingua, polo coñecemento e, en especial, pola ciencia.

Táboa de contidos

Resumo

Resumo

O carbono orgánico disolto (DOC) é o principal reservorio de materia orgánica disolta do océano, prodúcese de maneira autótrofa por organismos fotosintéticos e desempeña un papel fundamental no ciclo do carbono oceánico. Coñecer a súa distribución e o seu comportamento vólvese fundamental para comprender os posibles efectos do cambio climático sobre o océano global.

Porén, neste traballo, realízase o adestramento e implementación de distintos modelos de aprendizaxe automática co fin de identificar aquel que mellor prediga as concentracións do DOC no océano global a partir de variables físicoquímicas, a profundidade e o mes do ano. Os resultados indican que o modelo *Random Forest* ofrece o mellor axuste para a variable resposta considerada.

As predicións do DOC obtidas segundo o modelo axustado preséntanse nunha serie de climatoloxías e seccións verticais do océano para seren comparadas con outras estimacións atopadas na bibliografía. Estas arroxan bastante precisión nos resultados xunto coa novidosa integración da profundidade e da variación mensual nestas predicións, un aspecto chave neste traballo que abre un novo camiño neste campo de estudo.

English abstract

The Dissolved Organic Carbon (DOC) is the main reservoir of dissolved organic matter in the ocean; it is produced autotrophically by photosynthetic organisms and plays a key role in the ocean carbon cycle. Knowing its distribution and behaviour became essential to understand the possible effects of climate change over the global ocean.

Therefore, in this study, we fit and train different automatic learning methods to identify the best method that predicts the concentration of the global ocean DOC according to a set of physicochemical variables, depth and month of the year. The results prove that the *Random Forest* is the best method that suits our response variable.

Finally, the DOC predictions of this model are presented in a series of climatologies and vertical ocean sections to be compared with other estimations found in the literature. Our predictions demonstrate sufficient accuracy, together with the novel integration of the depth and monthly variations, revealing a key aspect of our research work that opens a new path in this field of study.

Capítulo 1

Introdución

Este primeiro capítulo ten como finalidade contextualizar e expor o problema abordado no presente traballo. En primeiro lugar, preséntase brevemente a entidade colaboradora na realización do mesmo: o CSIC-IIM. En segundo lugar, introdúcese ao lector no contexto do estudo, centrado nas dinámicas dos compoñentes mariños no océano global, cun enfoque específico no caso do carbono orgánico disolto (*Dissolved Organic Carbon*, DOC). Por último, explícase como o estudo e a mellora do coñecemento sobre a variabilidade do DOC, e o seu comportamento nos océanos, resulta esencial para avaliar os impactos potenciais do cambio climático.

1.1 CSIC-IIM e exposición do problema

O Instituto de Investigacións Mariñas (IIM), que forma parte do Consello Superior de Investigacións Científicas (CSIC), é un dos principais centros de investigación mariña da Península Ibérica. Este centro, situado en Vigo fronte ao Océano Atlántico, ten como obxectivo ampliar o coñecemento sobre o océano, dende as características físicas ata os aspectos relacionados coa cadea trófica, así como os procesos que interconectan os distintos compoñentes deste sistema.

No conxunto do sistema mariño, o presente traballo enfócase na ampliación de coñecementos sobre os procesos oceánicos asociados ao cambio climático. Unha maior comprensión dos procesos mariños é fundamental para entender, monitorizar e mitigar os efectos do cambio climático. Por este motivo, adóptase un enfoque multidisciplinar coa intención de enriquecer os estudos sobre o ciclo do carbono e, en particular, sobre o carbono orgánico disolto (DOC).

1.2 Carbono orgánico disolto

A materia orgánica nos océanos, que cobre o 71% da superficie terrestre (Thornton, 2014), adóitase definir como compostos actuando como doadores de electróns e protóns e, polo tanto, sendo unha fonte de enerxía metabólica para o crecemento dos organismos (Lonborg et al., 2020). Na construción destes compostos orgánicos xoga un papel fundamental a fotosíntese na superficie visible do océano, que contribúe a manter a cadea alimentaria nos ecosistemas mariños que presentan unha produción primaria estimada dun $40-50\%$ de toda a biosfera terrestre (Falkowski e Raven, 2013; Thornton, 2014). Este proceso de oxidación-redución (a fotosíntese) permite reducir químicamente o carbono mediante un aporte enerxético: a luz solar. A maioría de organismos fotosintéticos realizan este proceso nun contexto oxixenado, permitindo a liberación de osíxeno (O_2); entre eles atopamos as cianobacterias, as proclorófitas, as algas eucariotas e outras plantas superiores (Falkowski e Raven, 2013).

Habitualmente, na materia orgánica (OM) do sistema oceánico, realízase unha diferenciación funcional e arbitraria en dúas fraccións segundo a capacidade de filtración, definida polo tamaño de poro empregado (Verdugo et al., 2004; Thornton, 2014; Lonborg et al., 2020). Os materiais que quedan no filtro, cun tamaño de poro entre $0,2$ e $0,7$ μm, denomínanse materia orgánica particulada (*Particulated Organic Matter*, POM), mentres que os outros pertencen á categoría de materia orgánica disolta (*Dissolved Organic Matter*, DOM) (Verdugo et al., 2004; Lonborg et al., 2020). Ademais, inclúese unha distinción dentro da materia orgánica disolta (DOM) entre coloides[1] ou partículas submicroscópicas e verdadeiros solutos; ou incluso, noutros casos, unha división entre unha fracción de baixo peso molecular (*Low Molecular Weight*, LMW) e alto peso molecular (*High Molecular Weight*, HMW) (Verdugo et al., 2004).

Esta distinción operativa tamén ten algúns efectos bioxeoquímicos: a POM pode suspenderse na columna de auga ou almacenarse nos sedimentos, contén menor cantidade de biomasa no seu reservorio global e unha maior fracción de detritos[2], como células mortas; e inclúe a organismos como o zooplancto, fitoplancto ou bacterias. Por outra banda, o reservorio de DOM presenta unha menor cantidade de vida, aínda que presenta algúns organismos procariotas e virus, ten unha composición heteroxénea e transpórtase como soluto nas masas de auga oceánicas (Verdugo et al., 2004; Lonborg et al., 2020).

Na práctica, esta división da materia orgánica segundo o tamaño non se ve reflexada, senón que se observa un sistema continuo e dinámico. As moléculas orgánicas poden formar pequenas redes físicas autoensambladas máis estables por interaccións hidrofóbicas ou con certos ións, formando coloides macromoleculares e polímeros que, á súa vez, poden asociarse formando xeles[3] (nanoxeles e microxeles) (Verdugo et al., 2004; Hansell e Orellana, 2021). Debido ás interaccións e colisións moleculares, aínda poden formarse xeles de maior tamaño similares a partículas como as partículas exopoliméricas (*Transparent Exopolymer Particles*, TEP), que pertencen á POM (Verdugo et al., 2004).

Os sistemas mariños teñen unha concentración de DOM dun orde de magnitude superior á POM nas zonas fóticas[4] do océano, mentres que nos sedimentos a concentración de POM é tres veces maior. Ademais, a degradación biolóxica da DOM realízana, principalmente, os microorganismos; pola contra, son os organismos superiores os consumidores da POM. O reservorio desta materia orgánica disolta (DOM) é a mestura química máis complexa do planeta, onde só coñecemos unha parte dos posibles millóns de moléculas distintas que a conforman. Na parte coñecida do reservorio, as principais moléculas son: carbono, nitróxeno, fósforo, xofre, ferro e outros elementos esenciais para a vida; así como toxinas e contaminantes como o mercurio (Lonborg et al., 2020; Dittmar et al., 2021).

Centrándonos nas moléculas de carbono, o DOC é o maior reservorio de carbono reducido do océano, con máis de 200 veces o inventario de carbono da biomasa mariña (Hansell et al., 2009; Hansell, 2012; Lonborg et al., 2020; Hansell et al., 2024). A exportación e o seu consumo en augas profundas ten repercusións na retirada do carbono inorgánico disolto (*Dissolved Inorganic Carbon*, DIC) da superficie, que a súa vez modifica os intercambios de CO_2 entre o aire e o mar (Fontela et al., 2016). Porén, o DOC mariño xoga un importante papel na bioxeoquímica do ciclo de carbono oceánico (Carlson et al., 2010). Existe como almacén para o carbono producido, principalmente, de maneira autótrofa por organismos fotosintéticos como o plancto na superficie oceánica e emprégase como substrato para o metabolismo de organismos procariotas heterótrofos (Hansell et al., 2024).

A exportación deste DOC dende a superficie, por intercambio na columna de auga, ten un protagonismo fundamental para a bomba biolóxica, que consiste na suma de procesos que transportan o carbono bioxénico dende a superficie ata o interior dos océanos onde mineraliza. Esta exportación precisa dun gradiente vertical forte, dunha circulación activa mediada polo vento movendo as augas superficiais ricas en DOC e, eventualmente, dunha circulación que intercambie masas de auga en altas latitudes

[1] Un coloide é unha mestura non homoxénea de dúas ou máis fases: gas, líquido ou sólido.

[2] Os detritos son particulas residuais de tecidos, células ou materia orgánica liberada mediante procesos fisiolóxicos, patolóxicos ou de descomposición dos organismos.

[3] Os xeles consisten nunha rede tridimensional formada por polímeros que interactúan química e físicamente creando un pequeno entorno en equilibrio termodinámico con respecto ao medio oceánico que o rodea.

[4] A zona fótica é a rexión dos ecosistemas mariños onde penetra a luz solar.

e facilite os procesos de subdución nos xiros subtropicais. Ademais desta mestura vertical, tamén se produce a acumulación pasiva de partículas de carbono e o desprazamento vertical activo mediado polo zooplancton (Hansell et al., 2009; Carlson et al., 2010).

1.2.1 Concentración e distribución do DOC

As concentracións do DOC, a pesar de ser un gran repositorio, son relativamente baixas no océano aberto, oscilando nun rango entre os $\sim 35\ \mu molkg^{-1}$ das augas profundas ata os $\sim 70 - 80\ \mu molkg^{-1}$ superficiais presentes nos sistemas tropicais e subtropicais, onde existe unha maior estratificación vertical (Hansell et al., 2009; Hansell e Orellana, 2021). Como o fitoplancto é a principal fonte do DOM, e tamén do DOC, estas concentracións máis altas atópanse na capa produtiva superficial do océano (Dittmar et al., 2021). Ademais, de maneira habitual, obsérvanse aínda maiores concentracións nas augas costeiras que se atopan influenciadas pola desembocadura de sistemas fluviais (Hansell e Orellana, 2021).

A pesar destas baixas concentracións na maioría do océano ($\sim 40\ \mu molkg^{-1}$), o reservorio de DOC no océano divídese por un motivo operacional en distintas fraccións representadas na Táboa 1.1 (Hansell et al., 2009) e na Figura 1.1. Estas distintas tipoloxías existentes no conxunto do DOC nas augas oceánicas presentan características diferenciais entre elas, mais atópanse divididas principalmente pola súa labilidade biolóxica que consiste no tempo de vida ou persistencia no océano; e, ademais, presentan distintas funcións dentro do ciclo do carbono (Hansell, 2012; Hansell et al., 2024; Hansell e Orellana, 2021).

Fracción	**Inventario (Pg C)**	**Ratio de producción (Pg C / ano)**	**Ratio de retirada (µmol C / kg ano)**	**Vida útil (anos)**
Labile	$0,2$	~ 15 - 25	~ 10	$\sim 0,001$
Semi-labile	6 ± 2	$\sim 3,4$	~ 2 - 9	$\sim 1,5$
Semi-refractory	14 ± 2	$\sim 0,34$	$\sim 0,2$ - $0,9$	~ 20
Refractory	630 ± 32	$\sim 0,043$	$\sim 0,003$	~ 16.000
Ultra-refractory	> 12	$1,2 \times 10^{-5}$	8×10^{-7}	~ 40.000

Táboa 1.1: Táboa de fraccións do DOC no océano xunto con algunhas características cuantitativas de Hansell (2012).

Na parte máis superficial temos o *labile* DOC (LDOC): é a fracción que soporta a produción heterótrofa microbiana e ten unha baixa permanencia. Na zona fótica por debaixo da superficie temos o *semi-labile* DOC como aparece representado na Figura 1.1, sendo a fracción que máis contribúe á bomba biolóxica, e tamén o *semirefractory* DOC (SRDOC). Por último, a fracción do *refractory* DOC (RDOC) é bastante homoxénea en todo o océano e transpórtase de maneira conservativa pola circulación oceánica profunda (Hansell et al., 2009; Hansell, 2012; Hansell e Orellana, 2021).

Na distribución global deste conxunto de fraccións do DOC atópanse dous dominios ben diferenciados: o primeiro son as correntes superficiais impulsadas polo vento que chegan aos $\sim 200m$ de profundidade; e, o segundo, as masas de auga das profundidades oceánicas desprazadas polas correntes termohalinas (Hansell et al., 2009; Hansell e Orellana, 2021; Hansell et al., 2024). Estas correntes englóbanse dentro da circulación termohalina (*Meridional Overturning Circulation*, MOC), que aparece ilustrada na

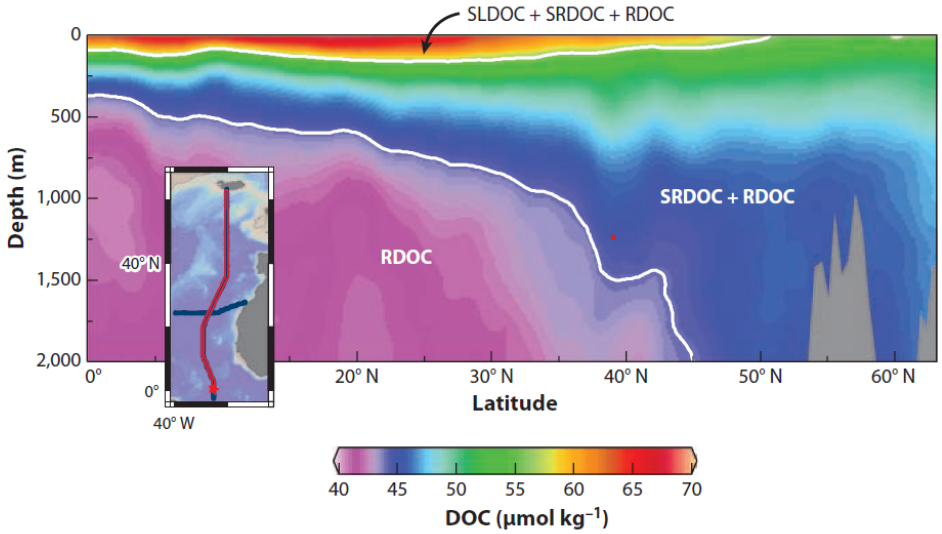

Figura 1.1: Figura representativa das distintas fraccións do DOC sobre unha sección vertical do Océano Atlántico, incluíndo un gradiente de cores relativo a súa concentración ($\mu molkg^{-1}$). Imaxe extraída de Hansell (2012).

Figura 1.2 e que consiste nunha circulación de intercambio pola cal as augas quentes superficiais que viaxan cara os polos, arrefrían e vólvense máis densas, polo que descenden e retornan cara as zonas ecuatoriais polas profundidades oceánicas (Toggweiler e Key, 2001).

Capas superficiais do océano

Na superficie oceánica é onde se acumula a maioría do novo DOC producido que escapa do consumo microbiano de organismos heterótrofos. As maiores concentracións de DOC ($\sim 70-80\ \mu molkg^{-1}$) lonxe dos continentes atópanse en baixas e medianas latitudes, asociadas aos xiros subtropicais das rexións ecuatoriais. Debido á estratificación por densidade pola incidencia solar, nestas zonas de maior salinidade, as capas superficiais enriquecidas en DOC e outros compoñentes permanecen sen mesturarse coas capas inferiores mediante mestura vertical (Hansell e Orellana, 2021; Hansell et al., 2024). Como excepción a isto, atópase o océano Ártico: moi estratificado e enriquecido en DOC procedente da materia orgánica liberada pola desembocadura de importantes ríos da zona ártica (Hansell et al., 2009, 2024).

Por outra banda, as menores concentracións na superficie atópanse no Océano Índico ($\sim 40-45\ \mu molkg^{-1}$) pola natureza diverxente das masas de auga, na que augas profundas con baixos niveis de DOC afloran nesa rexión. Ademais, en elevadas latitudes, a estabilidade vertical é menor, polo que a mestura convectiva pode ser máis profunda e o DOC producido na superficie reduce a súa concentración (Hansell e Orellana, 2021; Hansell et al., 2024).

De maneira excepcional, existen altas concentracións de DOC ($>\sim 80\ \mu molkg^{-1}$) no océano aberto. Algúns exemplos disto son os *hot spots* de DOC no Pacífico occidental ou os bordes das zonas continentais, onde presentan maiores concentracións asociadas á desembocadura de sistemas fluviais como no Amazonas, na costa este de Norteamérica ou no norte da baía de Bengala. Por último, mencionar algunhas rexións que teñen características diferenciais como o mar Negro, que recibe enormes cantidades de augas terrestres, polo que presenta concentracións superiores aos $150\mu mol/kg$; ou o caso

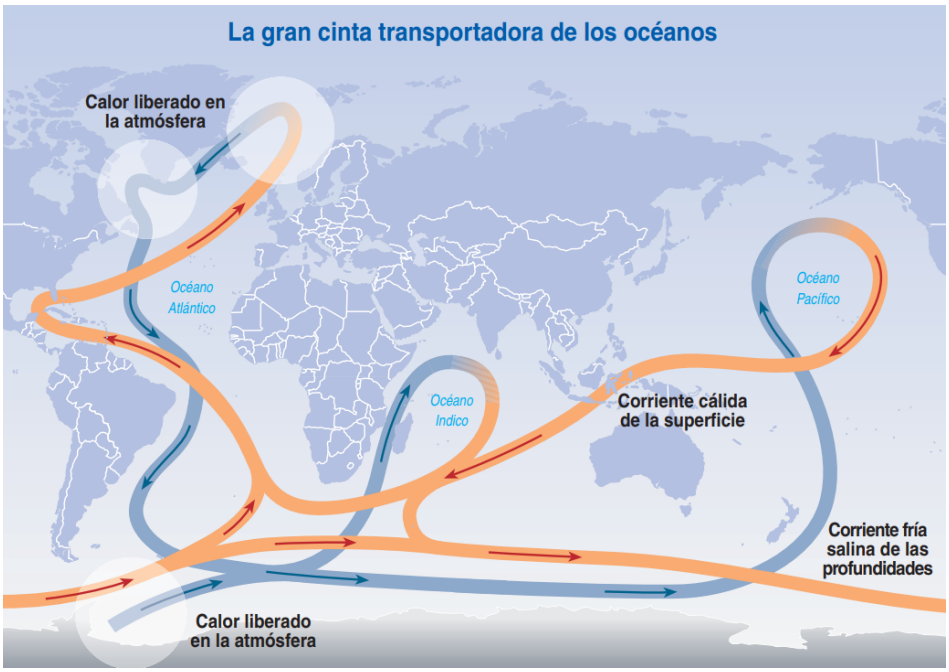

Figura 1.2: Mapa global da distribución da circulación termohalina (MOC). As seccións en laranxa representan o movemento de augas superficiais e, aquelas en azul, o transporte de augas profundas. Representación relativa ao informe de Watson (2001).

do mar Báltico, que tamén contén elevadas concentracións pola elevada produción autóctona en augas abertas e polas desembocaduras fluviais (Hansell et al., 2024).

Masas de auga profundas

Nas zonas fóticas dos xiros subtropicais, aquelas profundidades onde incide a luz solar, unha parte das augas enriquecidas en DOC pola actividade fotosintética pode ser exportada mediante o transporte de augas superficiais de Ekman[5] (Hansell et al., 2009). Pola contra, outra parte destas augas é transportada cara rexións oceánicas máis densas pola acción da corrente termohalina (MOC), onde arrefría e aumenta de densidade provocando o seu afundimento; como aparece ilustrado na Figura 1.2 (Hansell et al., 2009; Hansell e Orellana, 2021; Hansell et al., 2024).

O transporte máis importante de augas enriquecidas en DOC na circulación termohalina (MOC) acontece cara á rexión do Atlántico norte onde se produce a formación da corrente profunda NADW (Hansell et al., 2009; Carlson et al., 2010; Hansell et al., 2024), debido á mestura das augas superficiais subtropicais coas augas do Ártico (Bercovici et al., 2023). Nesta rexión, a entrada vertical de DOC produce un enriquecemento da zona batipeláxica (1.000-4.000 metros), cunhas concentracións entre os 45 $\mu mol/kg$ (Carlson et al., 2010; Hansell et al., 2024) ou incluso > 50 $\mu mol/kg$ (Hansell et al., 2009), como reflexa a Figura 1.3. A NADW marcha desta zona de formación no Atlántico Norte cara o sur guiada pola circulación termohalina (MOC) arredor dos 3.000 metros de profundidade, seguindo o patrón da *Deep Western Boundary Current* (DWBC) como remarca o artigo de Fontela et al. (2020).

Durante as correntes de ventilación das augas intermedias e profundas cara o Atlántico sur, a concentración do DOC vai decrecendo a medida que traspasa o ecuador a gran profundidade, chegando

[5]Movemento das masas de augas oceánicas con un certo ángulo con respecto á dirección do vento na capa superficial debido ao efecto de Coriolis.

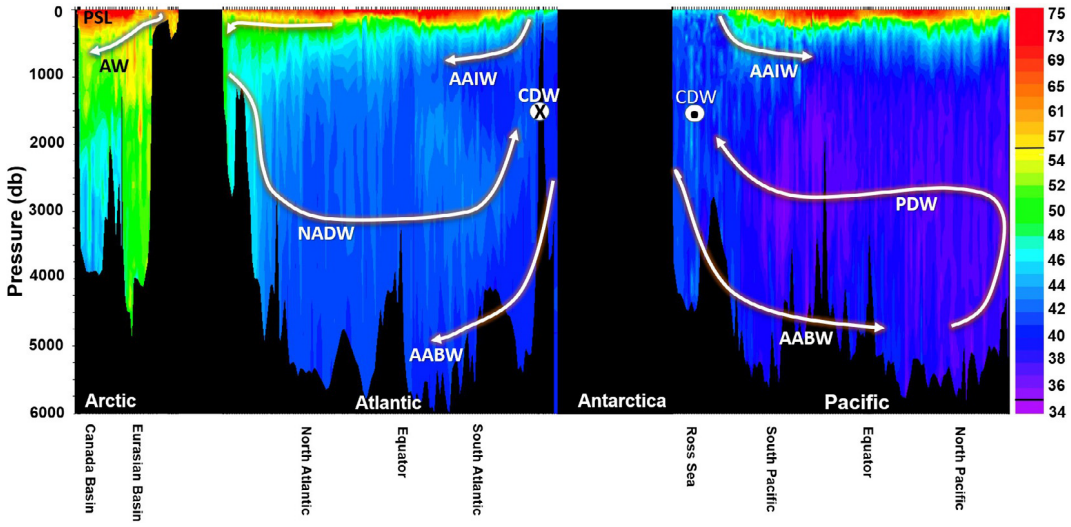

Figura 1.3: Sección meridional das concentracións de DOC ($\mu molkg^{-1}$) observadas ao longo de distintos océanos: Ártico, Atlántico, Antártico e Pacífico. Móstranse ademais as distintas masas de auga coas súas correntes asociadas, incluíndo a *Polar Surface Layer* (PSL), *Atlantic Water* (AW), *North Atlantic Deep Water* (NADW), *Circumpolar Deep Water* (CDW), *Antarctic Deep Water* (AABW), *Antartic Intermediate Water* (AAIW) e *Pacific Deep Water* (PDW). Figura procedente de Hansell et al. (2024).

ata valores de 39 $\mu molkg^{-1}$ no Océano Antártico profundo (Hansell et al., 2009). O destino final desta corrente é ser arrastrada pola CDW e desprazarse cara o Océano Índico (Bercovici et al., 2023; Hansell et al., 2024). Nestas rexións Antárticas, como no mar de Weddell, tamén atopamos a formación doutra corrente cara ás profundidades do océano: a AABW (Figura 1.3. Sen embargo, ao contrario que no Atlántico norte, non se observa o enriquecemento de DOC nas profundidades. Isto débese á existencia da Zona Frontal Polar do Océano Antártico que evita que o DOC enriquecido en baixas latitudes chegue a esta área de formación de augas profundas (Hansell e Orellana, 2021). A corrente NADW distínguese desta corrente AABW, a pesar de que ningunha está enriquecida en DOC, debido á diferenza de profundidades, maior temperatura e maiores salinidades (Bercovici et al., 2023; Hansell et al., 2024).

Estas augas profundas avanzan cara o Pacífico norte perdendo carbono a medida que se achegan ás zonas altas, dende os 42 $\mu molkg^{-1}$ ata os 36 $\mu molkg^{-1}$ cando chegan ás zonas profundas do norte (Hansell et al., 2009; Bercovici et al., 2023; Hansell et al., 2024). Esta perda durante o transporte débese a inxesta microbiana e a presenza dun importante sumidoiro de RDOC nas profundidades do Pacífico norte (Hansell et al., 2024). No instante que alcanzan estas rexións, prodúcese unha mestura vertical e un ascenso cara zonas intermedias onde se transforma na corrente PDW, que retorna cara o sur reducindo a concentración de DOC ata os 34 $\mu molkg^{-1}$ (Hansell et al., 2009; Bercovici et al., 2023; Hansell et al., 2024). Esta corrente PDW pode distinguirse claramente debido ao maior emprego de osíxeno aparente relativo (Bercovici et al., 2023).

1.2.2 Fontes e sumidoiros

Nos ecosistemas mariños, o DOC orixínase dende fontes autóctonas e alóctonas. No caso das autóctonas, os organismos fotosintéticos como o plancto son os principais produtores no océano (Hansell et al., 2009, 2024), sendo acompañados por algas e fluxos bentónicos e macrófitos nas zonas costeiras ou intermareais. Ademais, tamén se producen entradas alóctonas no sistema con orixe terrestre, como glaciares ou ríos, augas subterráneas e por intercambio coa atmosfera, como aparece subliñado na

Figura 1.4 (Lonborg et al., 2020).

Existen numerosos procesos polos que o fitoplancto libera materia orgánica como as infeccións virais líticas[6], a morte celular ou a exudación, chegando a liberar entre o 5% - 30% da súa produción primaria total. De maneira habitual, os mecanismos de liberación agrúpanse en filtracións: consisten nunha perda pasiva e exudacións; e transporte activo dende as células cara o exterior por exceso de carbono fixado durante a fotosíntese. Ámbolos dous mecanismos atópanse potenciados en condicións de elevada luminosidade e baixos nutrientes (Thornton, 2014; Lonborg et al., 2020).

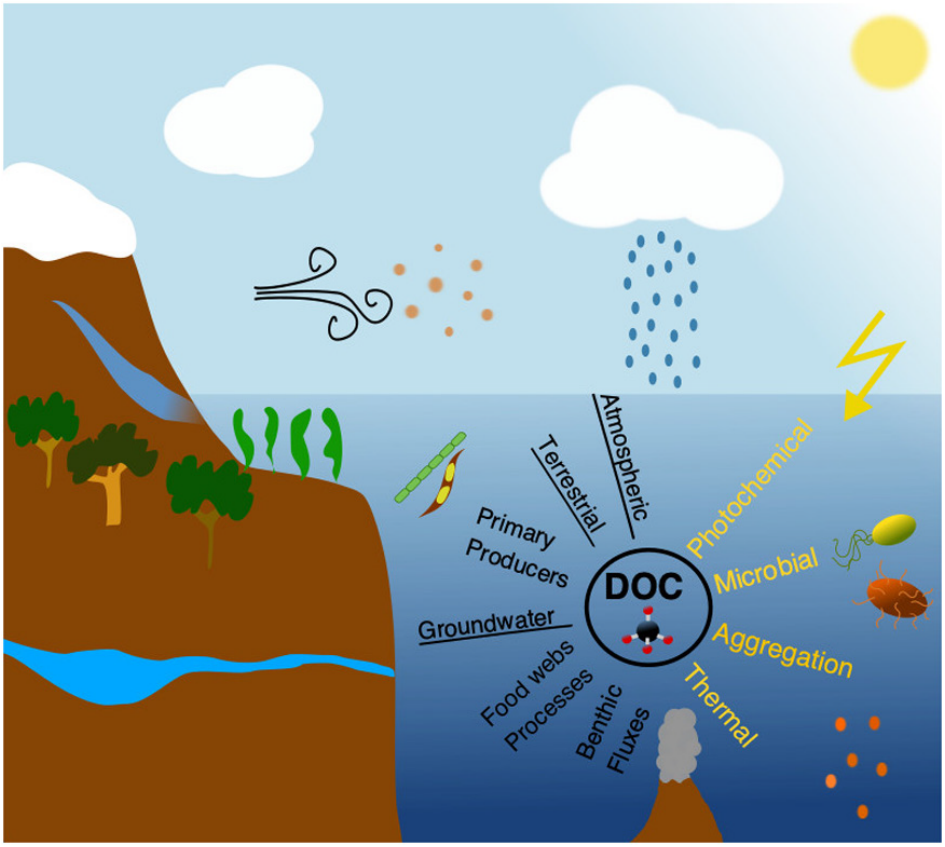

Figura 1.4: Resumo esquemático das principais fontes (texto en cor negra) e sumidoiros (texto en cor amarela) do reservorio de carbono orgánico disolto (DOC). Figura extraída de Lonborg et al. (2020).

Ademais desta liberación directa do DOC, tamén se produce un enriquecemento das augas profundas debido á disolución de partículas bioxénicas[7] afundidas, que forman parte do almacén de POM. Existen diversos procesos químicos e biolóxicos polos que se libera este DOC dende estas partículas afundidas, como poden ser a liberación pasiva, disgregación debido ás turbulencias ou disolucións enzimáticas. As partículas que afunden máis lentamente serven de substrato para as poboacións microbianas heterótrofas profundas, mentres que aquelas de maior densidade poden alcanzar rexións máis profundas (Hansell e Orellana, 2021; Hansell et al., 2024).

Este non é o único aporte nas profundidades oceánicas xa que outra das entradas importantes cara o océano son os fluxos bentónicos (Figura 1.4) dende os sedimentos mariños mediante un proceso de

[6] O ciclo viral lítico refírese ao proceso de infección dun virus no que cando penetra na célula, transcribe directa e rapidamente o material xenético e a célula morre por rotura ao liberarse as múltiples copias virais.

[7] Pertencente ou relativo á acción dos seres vivos.

degradación, sendo unha das maiores fontes de liberación do DOC comparable á entrada procedente dos sistemas fluviais. Por último, algúns estudos mostran que os sistemas xeotérmicos e as filtracións de petróleo tamén contribúen na liberación de DOC cara ás cuncas do océano profundo (Lonborg et al., 2020).

En canto os sumidoiros, o DOC producido pode ser rapidamente consumido por organismos heterótrofos (LDOC) ou pasar a formar parte do resto de fraccións coñecidas como DOC recalcitrante (Táboa 1.1). Este DOC recalcitrante refírese a aquelas moléculas que escapan desta degradación inicial, acumúlanse no reservorio de carbono e son observables no océano. Sen embargo, os heterótrofos, ademais de ser os principais consumidores, tamén contribúen, de maneira significativa, ao conxunto do DOC recalcitrante con aportacións de peptidoglicanos e outras macromoléculas nos procesos de división celular e lise viral (Hansell, 2012; Lonborg et al., 2020).

Existen diferentes procesos bióticos e abióticos responsables de retirar o DOC exportado ou recalcitrante, mais a súa importancia varía en función da fracción considerada (Hansell et al., 2009; Hansell, 2012). En términos xerais, os principais procesos de retirada de DOC da columna de auga do océano pódense englobar dentro dos mencionados na Figura 1.4 (Lonborg et al., 2020):

- Degradación termal: sistemas submarinos hidrotermais.
- Agregación: coagulación de burbullas e floculación abiótica cara micropartículas de maior tamaño ou sorción cara partículas.
- Reaccións fotoquímicas: degradación abiótica pola incidencia de luz.
- Heterótrofos: degradación biótica por organismos mariños.

Suxírese que a combinación dos efectos fotoquímicos e a degradación heterótrofa microbiana representa a maior parte das perdas no reservorio de DOC do océano (Lonborg et al., 2020). A incidencia de luz ultravioleta (UV) na superficie oceánica producindo fotólise, transformando e/ou interactuando coas partículas suspendidas, remata facilitando a remineralización microbiana (Hansell et al., 2009; Hansell, 2012). En zonas oceánicas máis profundas, a agregación e adsorción molecular do DOC cara o carbono orgánico particulado (*Particulate Organic Carbon*, POC) adquire unha maior relevancia, favorecida pola presenza de xeles orgánicos (Verdugo et al., 2004; Romera-Castillo et al., 2019; Hansell e Orellana, 2021).

Por outra banda, aínda que o consumo de O_2 está limitado nas zonas batipeláxicas (1.000-4.000 metros), a respiración total nestas zonas escuras segue sendo unha compoñente maioritaria dentro do fluxo de carbono; incluso a proporción de respiración procariota nas profundidades pode chegar a ser maior que na superficie (Hansell et al., 2009). O DOC convírtese nun elemento esencial neste fluxo e, porén, estímase que este reservorio xoga un papel imprescindible para manter a respiración heterótrofa nas profundidades do océano (Aristegui et al., 2002). Por último, en canto á relevancia, atópase a degradación termal: proceso hidrotermal a altas temperaturas onde as concentracións de DOC de saída son menores que na entrada (Lonborg et al., 2020).

1.3 Impacto do cambio climático

O principal obxectivo de incrementar o coñecemento sobre o ciclo de carbono e o seu funcionamento nos océanos é avaliar como pode verse afectado polo cambio climático e as posibles medidas para enfrontarse a este problema global. Segundo todo o mencionado ata agora, o reservorio de DOC mariño é un factor diferencial nos ecosistemas mariños, sendo a fonte de enerxía para os ecosistemas e o maior compoñente do ciclo do carbono terrestre (Lonborg et al., 2020). Polo tanto, debido ao continuo incremento da cantidade de CO_2 na atmosfera, o incremento da temperatura na superficie oceánica e ao descenso

do pH do océano (Thornton, 2014), vólvese imprescindible coñecer como estes cambios globais poden afectar ao ciclo do DOC oceánico.

Un dos principais factores estresantes do cambio global é o quentamento. Predícese que o océano, sendo o maior receptor do calor xerado polas emisións de gases de efecto invernadoiro, vai aumentar de media entre 1°C e 4°C a temperatura da superficie nas seguintes décadas (Lonborg et al., 2020). Este quecemento do océano vai provocar unha maior estratificación e, polo tanto, unha redución na mestura entre as augas superficiais e profundas (Thornton, 2014). Esta situación estresante de maior temperatura e menores nutrientes vai afectar ás comunidades de fitoplancto que reaccionan liberando maiores cantidades de materia orgánica disolta cara o entorno circundante (Engel et al., 2011; Thornton, 2014). A pesar desta maior liberación, a biomasa total dos produtores primarios vai reducirse e tamén se vai producir unha modificación na composición destas comunidades. En definitiva, este incremento da temperatura global vai afectar aos ratios e mecanismos de produción e degradación do DOC mariño (Lonborg et al., 2020).

Por outra banda, os niveis de CO_2 na atmosfera son un 40% superiores aos anteriores á Revolución Industrial debido ás actividades antropoxénicas. Este incremento vai afectar tanto incrementando as concentracións de DIC como modificando o equilibrio deste sistema, producindo á súa vez, un incremento das concentracións de CO_2 disolto e reducindo o pH do océano, nun proceso coñecido como acidificación do océano (Thornton, 2014; Lonborg et al., 2020). A pesar de que existen diversos estudos con resultados opostos e non existe un consenso unánime, parece que esta acidificación pode reducir os niveis de produción así como os de degradación microbiana do DOC, debido á conexión entre o intercambio de CO_2 na superficie oceánica e a exportación e consumo de DOC nas profundidades (Fontela et al., 2016; Lonborg et al., 2020). Por último, ademais dos dous procesos mencionados, existen outros mecanismos que poden afectar, tanto a curto como a longo prazo, ao ciclo do DOC oceánico e que se mencionan a continuación (Lonborg et al., 2020):

- Desoxixenación do océano
- Derretemento do xeo glaciar
- Incremento das entradas fluviais ao sistema
- Cambios nas correntes oceánicas
- Incremento dos afloramentos ou deposicións atmosféricas

Capítulo 2

Estado da arte

Debido ás implicacións e importancia do ciclo do carbono e, máis concretamente, do ciclo do DOC oceánico nos ecosistemas mariños, vólvese crucial comprender e predicir os posibles efectos que pode ter o cambio climático nestes sistemas e, por ende, no océano global. Neste sentido, diversas investigacións implementaron modelos estatísticos para realizar estimacións de diversas compoñentes do ciclo do carbono e, algunhas delas, foron enfocadas máis especificamente cara ao DOC. Porén, neste capítulo abordarase a explicación das distintas metodoloxías, bases de datos e aproximacións atopadas na literatura para exemplificar os traballos realizados ata agora neste contexto oceanográfico. Ademais, este resumo contribúe a contextualizar o obxectivo principal deste traballo: a predición da concentración do DOC ao longo de diferentes meses e a distintas profundidades do océano global, co fin de construír climatoloxías[1] representativas desta variable.

2.1 Ciclo do carbono

A importancia do ciclo do carbono na xeodinámica terrestre e a súa posible modificación debido ao cambio climático, facilitou que diversos investigadores de xeito paralelo comezasen a traballar na estimación de diversas compoñentes deste ciclo. De maneira concreta, pódese observar a Táboa 2.1, onde aparecen recollidos diversos artigos onde implementan diversas metodoloxías para conseguir estas predicións.

En primeiro lugar, no artigo de Liu et al. (2021) estiman as concentracións de POC na superficie do océano. En concreto, os autores realizan unha procura inicial para atopar os puntos coincidentes entre os datos recollidos *in situ* sobre medicións do POC procedentes de dúas bases de datos e os datos satelitais procedentes da Axencia Espacial Europea (*European Space Agency*, ESA). Estes datos satelitais consisten en datos procedentes da teledetección do color do océano. A continuación, empregan os datos coincidentes na mesma posición da superficie do océano, unha mostra de 3.551 observacións, para adestrar e validar tres tipos de modelos: un *Extreme Gradient Boosting* (XGBoost) (Chen and Guestrin, 2016), un *Support Sector Machine* (SVM) (Drucker et al., 1996) e unha rede neuronal artificial (*Artificial Neural Network*, ANN) (Specht, 1991).

Dun xeito semellante, na investigación realizada por Wu et al. (2023) empréganse datos *in situ* procedentes de tres bases de datos e datos satelitales do color do océano pertencentes á NASA (*National Aeronautics and Space Administration*). Despois de realizar unha procura de puntos coincidentes, obtéñense un total de 14.017 observacións no período entre 2007 e 2017 que van ser empregadas para adestrar, validar e probar os modelos axustados. Ademais, neste caso, faise unha distinción en tres tipos

[1] As climatoloxías son representacións visuais da superficie terrestre onde se mostra a distribución de certa variable físicoquímica nos diferentes océanos do planeta.

de augas oceánicas en función da porcentaxe de POC con respecto á materia particulada suspendida (*suspended particulate matter*, SPM). Nesta aproximación formuláronse ata seis modelos distintos: rede neuronal convolucional (*convolutional neural network*, CNN) (Lecun et al., 2002), rede neuronal retropropagada (*backpropagation neural network*, BPNN)(McCulloch e Pitts, 1990; Lecun et al., 2015), redes neuronais basadas na atención (*attention-based neural network*, ABNN) (Yang et al., 2019), *Adaptative Boosting* (AdaBoost) (Freund e Schapire, 1995), *Extreme Gradient Boosting* (XGBoost) e *Random Forest* (Breiman, 2001). Por último, é importante destacar o emprego de ata 30 covariables co obxectivo de estimar as concentracións de POC na superficie global durante o período temporal 2007-2017 e, logo, comparalas cos datos de POC procedentes da base de datos da NASA.

Ademais, existe outra investigación de Li et al. (2024) na que tamén se empregan diversos modelos de aprendizaxe automática pero restrinxíndose ao contexto do mar Mediterráneo. Neste traballo empregan ata un total de 47 variables ambientais procedentes de datos satelitais para predicir o POC da superficie do Mediterráneo, empregando medidas *in situ* para validar as estimacións. Os modelos empregados por estes autores foron: rede neuronal retropropagada (BPNN), *AdaBoost* (Ying et al., 2013), *Extreme Gradient Boosting* e *Random Forest*.

Artigo	Obxectivo	Bases de datos	Metodoloxía
Liu et al. (2021)	Predición POC superficial	Datos *in situ* POC e datos satelitais do color do océano	Modelo *XGBoost*, SVM e ANN
Wu et al. (2023)	Predición POC superficial	Datos *in situ* POC e datos satelitais do color do océano	Modelo CNN, BPNN, ABNN, *XGBoost*, *AdaBoost* e *Random Forest*
Li et al. (2024)	Predición POC Mediterráneo	Datos *in situ* e satelitais no Mediterráneo	Modelo BPNN, *AdaBoost*, *XGBoost* e *Random Forest*
Lee et al. (2019)	Predición TOC solo mariño	Observacións empíricas TOC no solo mariño	Modelo kNN
Jin et al. (2024)	Predición PIC e POC	Datos satelitales PIC e POC e WOA 2018	Modelo *Random Forest*
Broullón et al. (2020)	Predición TCO_2 mensual	Bases de datos empíricas globais	Modelo *feedforward neural network*

Táboa 2.1: Resumo descritivo dos distintos artigos que empregan modelos estatísticos para estimar diversas compoñentes do ciclo do carbono nos océanos.

Por outra banda, tamén se atopan estimacións enfocadas noutras seccións do ciclo do carbono, como é o caso do artigo de Lee et al. (2019), que se centra en predicir o carbono orgánico total (*Total Organic Carbon*, TOC) do solo mariño empregando o modelo dos k-veciños máis próximos (*k-nearest neighbors*, kNN) (Zhang, 2016) ou, a investigación de Jin et al. (2024), onde empregan un *Random Forest* para comparar a sensibilidade do carbono inorgánico particulado (*Particulate Inorganic Carbon*, PIC) e o POC a diferentes condicións ambientais. Por último, é importante incluír o traballo realizado por Broullón et al. (2020) onde abordan a problemática de estimar mensualmente o carbono disolto inorgánico total (*total dissolve inorganic carbon*, TCO_2) mediante un algoritmo de redes neuronais

retroalimentadas (*feedforward neural network*) (Rumelhart et al., 1986).

2.2 Predición DOC

A problemática da estimación e predición das concentracións do DOC no océano global a partir doutras variables oceánicas é menos popular, sendo un campo de estudo aberto e con moitas posibilidades. Existen moi poucos estudos na bibliografía que o aborden e, aqueles artigos que o fan, atópanse resumidos na Táboa 2.2.

Por un lado, temos o estudo de Roshan e DeVries (2017) onde realizan unha avaliación a escala global da produción e exportación do DOC. Empregan unha combinación de redes neuronais artificiais (ANN), estimadas en función da distribución global do DOC, xunto cun modelo de circulación oceánica. No axuste e adestramento deste modelo foron empregadas arredor de 30.000 observacións, procedentes de diversas bases de datos, que conteñen a variable do DOC xunto con outras como os nitratos, profundidade, salinidade, etc., todas elas situadas no mesmo patrón oceánico de 1° de lonxitude por 1° de latitude, cubrindo todo o océano. Unha vez estimados os valores da concentración do DOC no océano global, posteriormente, realizáronse cálculos sobre a produción e exportación; relacionándoas coa produción de nitratos nos océanos oligotróficos. Outra das cuestións diferenciais neste artigo é a estimación do DOC en distintas profundidades: 20 metros, 300 metros e 600 metros; convertíndose na primeira metodoloxía que diferencia as estimacións segundo a compoñente espacial da profundidade.

Ademais, tamén se atopa a recente investigación realizada por Bonelli et al. (2022) onde empregan algoritmos de redes neuronais artificiais (ANN) para predicir as concentracións de DOC a partir de datos físicoquímicos, concentracións de clorofila e valores do coeficiente de absorción da materia orgánica disolta coloreada. Neste caso, o proceso seméllase moito ao empregado nos artigos referidos previamente, como o de Liu et al. (2021) ou o de Wu et al. (2023), mais neste caso aplicado á estimación do DOC. Empréganse datos *in situ* e datos satelitales para realizar unha procura dos puntos superficiais coincidentes e así empregar estes datos no axuste e validación do modelo de ANN. Esta aproximación difire da proposta neste traballo e, aínda que inclúe diversas variables físicoquímicas, non engade a compoñente temporal ou espacial da profundidade na construción dos modelos.

Artigo	Obxectivo	Bases de datos	Metodoloxía
Roshan e DeVries (2017)	Predición da produción e exportación do DOC	Diferentes bases de datos empíricas	Modelo ANN
Bonelli et al. (2022)	Predición DOC superficial	Bases de datos *in situ* e satelitales	Modelo ANN
Laine et al. (2024)	Predición DOC superficial e anual	Bases de datos *in situ* e satelitales	Modelo lineal, *Gradient Boosting* e *Random Forest*
Panaïotis et al. (2025)	Predición DOC en profundidade	Datos de Hansell et al. (2021)	Modelos BRTs

Táboa 2.2: Resumo descritivo daqueles artigos que empregan modelos estatísticos para estimar o carbono orgánico disolto (DOC) no océano global.

Recentemente, Laine et al. (2024) estiman as concentracións do DOC no océano global a partir de datos

satelitais en combinación con observacións *in situ*. Estes autores manexan datos de absorción dos océanos en seis diferentes lonxitudes de onda xunto con outras variables como a produción primaria, a salinidade e a temperatura da superficie para estimar os valores do DOC. Para o proceso de adestramento dos modelos propostos utilizan datos globais do DOC observados *in situ*, en diferentes campañas oceanográficas procedentes da mesma base de datos deste traballo: Hansell et al. (2021). Despois da selección dos datos superficiais e dunha procura de coincidencias en localización e temporalidade cos datos satelitais, obtéñense 8.796 observacións. No canto de traballar coa mesma resolución espacial e temporal ca os datos satelitais, finalmente trabállase con 1.339 mostras cun total de 13 potenciais variables explicativas para predicir os valores de DOC. Estas variables foron as seis lonxitudes de onda, temperatura, latitude, profundidade, produción primaria, salinidade e distancia coa costa. Unha vez preparada a base de datos, axustan os diferentes modelos: modelo lineal múltiple, *Random Forest* e *Gradient Boosting*. O *Random Forest* obtivo os mellores resultados na predición do DOC para o océano aberto, xerando como resultado unha serie de climatoloxías mensuais da superficie do océano durante o período 2010-2018. Sen embargo, a pesar de incluír a profundidade como variable preditora, non se realizan climatoloxías en función da compoñente espacial da profundidade, só en superficie.

Por último, no recente traballo de Panaïotis et al. (2025) abórdase a construción de climatoloxías de predición do DOC en catro niveis de profundidade diferentes, empregando para isto, o método de aprendizaxe automática dos *Boosted Regression Trees* (BRTs) que combinan as árbores de regresión (Breiman et al., 1984) coas propiedades dos procesos *boosting* (Schapire, 2003). Neste traballo utilízase a base de datos de Hansell et al. (2021) para a construción do modelo, eliminando aquelas observacións problemáticas e outras procedentes de fontes autóctonas. Estas observacións do DOC complementáronse cunha serie de preditores ambientais procedentes de climatoloxías mensuais do *World Ocean Atlas* de 2018 (WOA18), sobre un patrón de 1°de lonxitude por 1°de latitude para cubrir todo o océano global. Os autores aplicaron os modelos BRTs para predicir as concentracións do DOC en función das variables bioxeoquímicas engadidas, combinando nalgunhas ocasións as variables da profundidade na que se realiza a predición coas variables das capas superiores. No proceso de adestramento e procura de hiperparámetros utilizouse a validación cruzada para, finalmente, construír as climatoloxías nas diferentes rexións de profundidade do océano. Estas seccións onde se realizan as climatoloxías son: superficie ($0 - 10$ metros), capa epipeláxica ($10 - 200$ metros), capa mesopeláxica ($200 - 1000$ metros) e capa batipeláxica (> 1000 metros).

A partir da bibliografía revisada, o principal obxectivo deste traballo é contribuír e ampliar os coñecementos existentes no eido da modelaxe do carbono orgánico disolto (DOC), resumidos ao longo deste capítulo, no marco da investigación oceanográfica. Para acadar este propósito, proponse a implementación de novos modelos de aprendizaxe automática e o emprego de bases de datos máis completas. Como xa foi comentado, a finalidade última consiste en predicir a distribución do DOC no océano global a partir de distintas variables explicativas de natureza físicoquímica das augas oceánicas, co obxectivo de xerar climatoloxías representativas para os diferentes meses do ano e para as diversas profundidades consideradas.

Aínda que esta aproximación non resulta completamente innovadora —xa que as climatoloxías mensuais foron previamente abordadas en Laine et al. (2024), e a estratificación por profundidades en Panaïotis et al. (2025)—, o presente traballo pode converterse nun estudo moito máis completo e integrador. A metodoloxía propón incluír simultaneamente a dimensión temporal, mediante a variable mes,e un número significativamente maior de niveis de profundidade na predición das concentracións do DOC, co obxectivo de construír climatoloxías a escala global. Polo tanto, este enfoque ofrece unha perspectiva diferencial sobre a distribución do DOC, a súa relevancia dentro do ciclo bioxeoquímico do carbono oceánico e, en última instancia, o seu papel fronte ao cambio climático.

Capítulo 3

Bases de datos

Neste capítulo procédese a presentar, explicar e realizar unha análise exploratoria das bases de datos empregadas para estudar e aplicar os modelos estatísticos para a estimación do DOC, que serán presentados no seguinte capítulo. En primeiro lugar, emprégase a base de datos de Hansell et al. (2021) para adestrar e validar os modelos. A continuación, para predicir as concentracións da nosa variable de estudo (DOC) en todo o océano global, utilízanse os datos procedentes do *World Ocean Atlas* (WOA) do ano 2023 (Reagan et al., 2024).

3.1 Base de datos de Hansell

Este conxunto de datos consiste nunha serie de medidas da materia orgánica disolta (DOM) e outros parámetros químicos e hidrográficos procedentes de observacións oceánicas globais obtidas no período entre 1994 e 2021. O obxectivo principal desta recompilación de datos procedentes de mostras de botellas de barcos é crear unha colección accesible, directa e uniforme dos datos relacionados coa DOM recollidos por diferentes laboratorios e investigadores ao longo de todo o mundo. Máis concretamente, nesta colección hai 268.325 observacións procedentes de 230 travesías científicas enfocadas principalmente no DOC e coa intención de cubrir todo o océano global. Conteñen un total de 96 variables: hidrográficas (ex: temperatura, salinidade), bioxeoquímicas (ex: osíxeno, nutrientes), procedentes do sistema do carbono inorgánico (ex: DOC, pH ou alcalinidade) e biolóxicas (ex: clorofila), entre outras.

Antes de comezar a traballar sobre esta base de datos, realízase un procesado inicial dos datos onde se seleccionan as variables de interese, elimínanse os valores negativos e suprímense as observacións nulas ou ausentes. Neste traballo, as variables explicativas escollidas atópanse restrinxidas pola presenza na base de datos WOA 2023, que vai ser o conxunto de observacións empregado para realizar as predicións das concentracións da nosa variable DOC. Aquelas variables que non están en WOA 2023 van ser eliminadas.

Despois deste proceso inicial, temos un total de 72.964 observacións con 11 variables distintas distribuídas nos puntos de mostra reflectidos no mapa do mundo representado na Figura 3.1. Ademais das compoñentes verticais e horizontais destes puntos de observación, correspondentes á latitude (`LATITUDE`) e lonxitude (`LONXITUDE`), tamén imos empregar a compoñente da profundidade que non aparece nesta figura.

Estes datos foron descargados dende a seguinte páxina web: base de datos Hansell (2021).

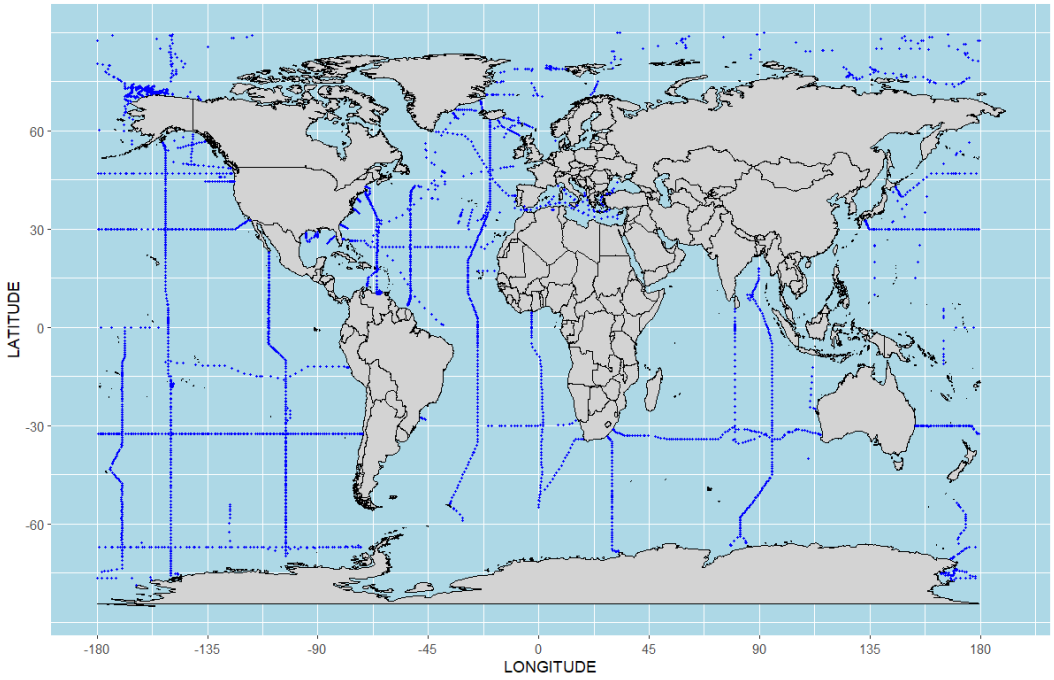

Figura 3.1: Mapa do mundo onde se representan os puntos de mostraxe da base de datos de Hansell et al. (2021) que imos empregar no adestramento dos modelos.

3.2 Base de datos de WOA

Este atlas consiste nunha colección de medias interpoladas polo océano global de distintas variables físicoquímicas, distintas profundidades analizadas, distintos períodos de tempo e incluso distintas resolucións ás que foron observados e calculados estes datos (Reagan et al., 2024). Porén, á hora de utilizar esta base de datos, foi necesaria unha selección previa dos arquivos necesarios para completar as predicións dos modelos. A resolución dos datos empregada foi un patrón de tamaño 1°de lonxitude por 1°de latitude, que contén algo máis de 70.000 observacións na maioría das profundidades.

Como resumo descritivo dos arquivos seleccionados, mostrando a profundidade e períodos temporais escollidos, preséntase a Táboa 3.1. En todas as variables emprégase a media obxectiva analizada. A intención neste estudo é escoller as observacións máis recentes e completas coa resolución do patrón oceánico 1°de lonxitude por 1°de latitude; sen embargo, non é posible en todos os casos. Deste xeito, como aparece nesa táboa, para cada unha das variables físicoquímicas escóllese o intervalo temporal máis completo, seleccionando todas as profundidades que conteñen datos medidos de maneira mensual. Nas primeiras tres variables (temperatura, salinidade e osíxeno) ata os 1.500 metros e, nas tres seguintes (nitratos, fosfatos e silicatos), ata os 500 metros.

A continuación, para obter os datos completos ata os 5.500 metros de profundidade en todas as variables, escóllense arquivos de datos estacionais ou anuais. No caso dos arquivos estacionais, os meses correspondentes á mesma estación obteñen os mesmos valores dende os 1.550 metros ata os 5.500 metros nas dúas primeiras variables (temperatura e salinidade). E, para o resto de variables, todos os meses do ano presentan os mesmos valores nas profundidades indicadas na última columna da Táboa 3.1. Deste xeito, asumindo certas condicións, garántese a existencia dunha base de datos completa con todas as variables de interese e todas as profundidades escollidas e sobre unha malla de 1°de resolución sobre o océano global.

Estes datos foron descargados dende a seguinte páxina web: *WOA data*.

Variables	Dim. temporal	Prof. mensuais	Prof. estacionais	Prof. anuais
Temperatura	1991-2020	0-1.500m	1.550-5.500m	-
Salinidade	1991-2020	0-1.500m	1.550-5.500m	-
Osíxeno	1991-2020	0-1.500m	-	1.550-5.500m
Nitratos	1965-2020	0-500m	-	550-5.500m
Fosfatos	1965-2020	0-500m	-	550-5.500m
Silicatos	1965-2020	0-500m	-	550-5.500m

Táboa 3.1: Resumo das profundidades empregadas para obter os datos da media obxectiva analizada pada unha das variables cunha resolución de 1°por 1°. Indícase a dimensión temporal destas observacións ("Dim. temporal") xunto co tipo de medición temporal escollida en función de cada profundidade: mensual ("Prof. mensuais"), estacional ("Prof. estacionais") ou anual ("Prof. anuais").

Despois do proceso de descarga, axuste e unión dos distintos arquivos correspondentes, xurde un conxunto de datos formado por 19.722.240 observacións con 10 variables preditoras: temperatura, salinidade, osíxeno, nitratos, fosfatos, silicatos, latitude, lonxitude, profundidade e mes. En concreto, na variable da profundidade incluíronse un total de 40 niveis de profundidade oceánica: 0, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 125, 150, 175, 200, 225, 250, 275, 300, 350, 400, 450, 500, 600, 700, 800, 900, 1.000, 1.250, 1.500, 1.750, 2.000, 2.500, 3.000, 3.500, 4.000, 4.500, 5.000 e 5.500 metros. Todas elas ao longo dos 12 meses do ano. A partir desta extensa base de datos vanse realizar as predicións da nosa variable resposta, a concentración do DOC.

3.3 Análise exploratoria de datos

No proceso de exploración dunha base de datos, cando existe un gran número de variables explicativas que poden ou non ser relevantes para predicir a variable resposta, resulta útil reducir o modelo matemático empregando algún método de selección de variables. Mais, neste traballo, as variables explicativas empregadas atópanse limitadas pola base de datos WOA 2023, sobre a que imos realizar as predicións da variable resposta de interese (DOC). Porén, aquí se expón unicamente un resumo descritivo das variables que se incluirán nos modelos seguindo as indicacións de nomenclatura de datos oceanográficos recollidas no artigo de Jiang et al. (2022):

- LATITUDE: latitude (grados decimais). Latitude medida en grados decimais norte.
- LONGITUDE: lonxitude (grados decimais). Lonxitude medida en grados decimais este.
- DEPTH: profundidade (metros). Profundidade en metros á que a mostra é recollida
- CTDTMP: temperatura (grados celsius). Temperatura medida *in situ* mediante CTD seguindo a escala ITS-90.
- CTDSAL: salinidade (sen unidades). Salinidade calculada pola condutividade medida con CTD empregando a ecuación da Escala Práctica de Salinidade de 1978 (PSS-78).
- CTDOXY: osíxeno disolto ($\mu molkg^{-1}$). Contido de osíxeno disolto (O_2) medido por sensores de osíxeno contidos no CTD.

- SILCAT: silicatos ($\mu molkg^{-1}$). Contido total de silicatos inorgánicos disoltos: $Si(OH)_4$, H_4SiO_4, SiO_2, Sil.
- NITRAT: nitratos ($\mu molkg^{-1}$). Contido de nitratos (NO_3^-).
- PHSPHT: fosfatos ($\mu molkg^{-1}$). Contido total de fosfatos inorgánicos disoltos: $H_2PO_4^-$, HPO_4^{2-}, PO_4^{3-}
- MONTH: mes. Mes do calendario en formato UTC onde son rexistrada cada unha das mostras.

Ademais das distintas variables explicativas, tense a variable resposta:

- DOC: carbono orgánico disolto ($\mu molkg^{-1}$). Contido total de carbono orgánico disolto.

Na Táboa 3.2 descríbense as diferentes variables coas respectivas unidades de medida e unha serie de medidas de centralidade como a media e a mediana xunto cos valores mínimo e máximo, sendo unha aproximación do rango de valores da distribución de cada unha delas. Primeiro, aparecen variables espaciais como a `LATITUDE` e `LONGITUDE` cun rango de valores acorde ao mapa mundial. Deste xeito, a latitude debe atoparse entre os 90°N e os −90°N e, a lonxitude, entre os 180°E e −180°E. Neste caso, as medidas de centralidade non teñen excesiva repercusión nestas variables de posición. Ademais, tamén se atopa a variable profundidade (`DEPTH`) cun rango de valores entre 0 metros e 8.556 metros, cun valor medio de aproximadamente 4.000 metros.

Variable	**Unidades**	**Mínimo**	**Mediana**	**Media**	**Máximo**
LATITUDE	grados norte	−78,64	2,00	0,198	89,988
LONGITUDE	grados este	−179,99	−52,36	−34,06	179,85
DEPTH	metros	9,00	4.292,00	4.001,00	8.556,00
CTDTMP	grados Celsius	−2,083	3,867	7,052	31,221
CTDSAL	sen unidades	17,53	34,70	34,72	39,25
CTDOXY	$\mu molkg^{-1}$	0,00	200,40	189,20	649,20
SILCAT	$\mu molkg^{-1}$	0,00	29,23	53,38	231,13
NITRAT	$\mu molkg^{-1}$	0,00	25,31	22,65	47,16
PHSPHT	$\mu molkg^{-1}$	0,000	1,790	1,626	6,550
MONTH	sen unidades	1,000	4,000	5,329	12,000
DOC	$\mu molkg^{-1}$	27,50	43,09	47,30	149,28

Táboa 3.2: Táboa descritiva das variables pertencentes á base de datos de Hansell et al. (2021).

Por outra banda, pódense observar as variables físicoquímicas, con valores superiores a 0 debido ao filtrado inicial da base de datos; exceptuando a temperatura (`CTDTMP`), onde o valor mínimo é de −2.083°C. Algunhas variables teñen unha distribución moi ampla como o osíxeno disolto (`CTDOXY`): con valores dende os 0 $\mu molkg^{-1}$ ata os 649,2 $\mu molkg^{-1}$; ou os silicatos (`SILCAT`), de 0 a 231,13 $\mu molkg^{-1}$. Pola contra, outras presentan menores diferenzas entre os valores mínimos e máximos como, por exemplo, a salinidade (`CTDSAL`): dende os 17,53 ata os 39,25; ou o fosfato (`PHSPHT`): de

$0,000$ a $6,550$ $\mu molkg^{-1}$. Por último, aparece a variable mensual (`MONTH`), sendo o valor mínimo 1 que corresponde co mes de xaneiro, mentres que o valor máximo 12 co correspondente mes de decembro.

Ao marxe das variables explicativas, na mesma Táboa 3.2, atópase a variable resposta `DOC` cuns valores mínimos nas profundidades do océano de $27,50$ $\mu molkg^{-1}$ e unha concentración máxima restrinxida aos 150 $\mu molkg^{-1}$ durante o procesamento inicial, seguindo as indicacións sobre a súa distribución e concentración explicadas durante a introdución deste traballo. Destaca tamén que o valor medio da concentración desta variable é de $47,30$ $\mu molkg^{-1}$ e a mediana de $43,09$ $\mu molkg^{-1}$, valores próximos ás concentracións atopadas na maioría do océano descritas por Hansell et al. (2009).

Despois de realizar unha análise descritiva inicial do conxunto de variables da base de datos, calcúlanse as correlacións entre as distintas variables co obxectivo de observar as relacións existentes e discernir o seu comportamento. Na Figura 3.2 móstrase o gráfico de correlacións entre as distintas variables, onde unha elevada correlación positiva, próxima a 1, aparece coa cor azul escura e cunha elipse aproximándose a unha liña recta; mentres que cando hai unha elevada correlación negativa, aproxímase a -1 e recibe unha cor vermella escura. Ademais, cando existe pouca correlación e os valores tenden a ser menores a $0,5$ ou próximos a 0, a figura elíptica debuxada aproxímase cara a un círculo.

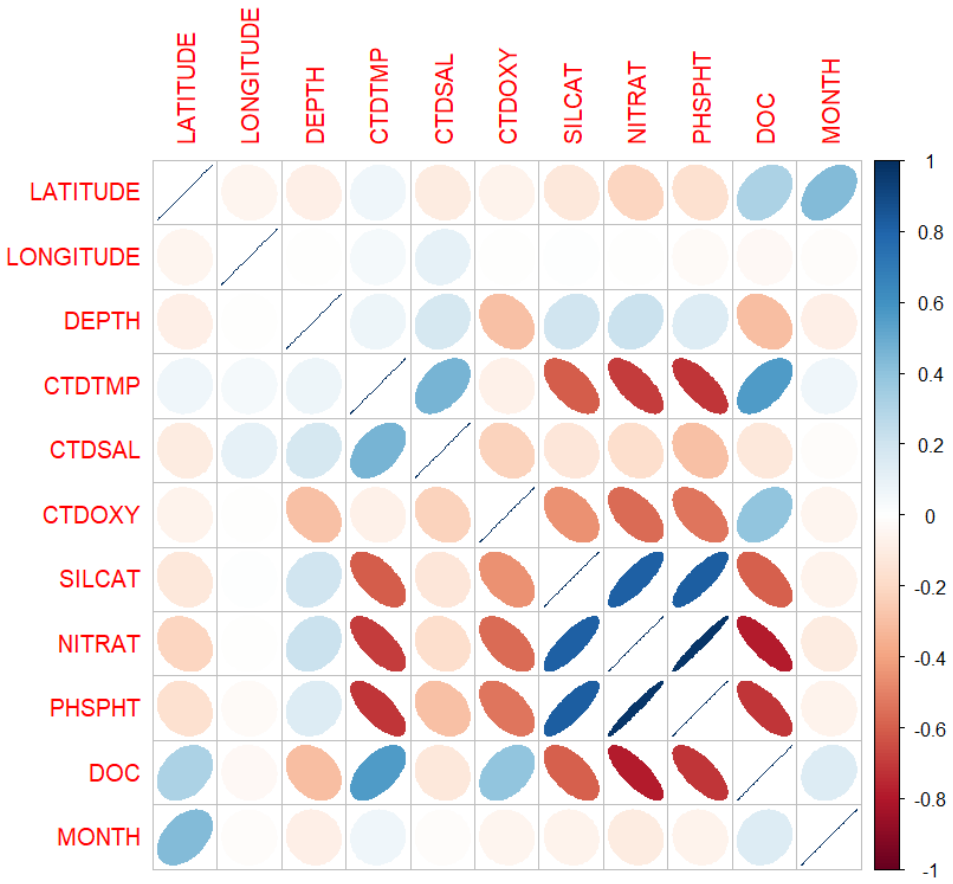

Figura 3.2: Gráfico de correlacións entre as distintas variables da base de datos de Hansell et al. (2021).

Deste xeito, obsérvase como as variables `SILCAT`, `NITRAT` e `PHSPHT` teñen unha elevada correlación positiva entre elas, cunha cor vermella e figura elíptica. Ademais, estas tres variables presentan unha correlación negativa relativamente alta coa temperatura (`CTDTMP`) ou co `DOC`, e algo menor co osíxeno

disolto (`CTDOXY`). Tamén se pode destacar certa correlación positiva da variable `DOC` coa temperatura (`CTDTMP`), aínda que en menor medida. Para rematar, mencionar que o resto de correlacións entre variables teñen valores baixos e moi pouco salientables.

No canto de coñecer un pouco mellor a distribución das variables, ao marxe das métricas mencionadas na Táboa 3.2, realízase un gráfico *box-plot* presentado na Figura 3.3. Nesta Figura descríbense os gráficos *box-plot* correspondentes a seis variables físicoquímicas de interese. Nesta descrición suprimíronse as variables de lonxitude, latitude, profundidade e mes. Pode observarse que a variable `CTDSAL` contén a maioría dos datos arredor do valor 35, e presenta unha gran parte dos datos atípicos. Os datos desta variable concéntranse moito nun rango de valores moi reducido, outorgándolle unha distribución neste gráfico bastante complexa e moi pouco visual. Pola contra, tamén temos outras variables onde non se aprecian datos atípicos como os silicatos (`SILCAT`) ou nitratos (`NITRAT`). Os primeiros presentan unha mediana bastante desprazada cara valores baixos, mentres que a outra variable mostra unha mediana superior.

Ademais, tamén observamos como as variables de temperatura (`CTDTMP`) e fosfatos (`PHSPHT`) agrúpanse en torno a valores reducidos, con mediana de 5°C e preto dos 2 $\mu molkg^{-1}$, respectivamente. Tendo, ámbalas dúas, un conxunto de valores atípicos na parte superior. Por último, no caso da variable do osíxeno disolto (`CTDOXY`), obsérvase que os valores están bastante agrupados en torno á mediana, cun rango de valores relativamente amplo ata os 400 $\mu molkg^{-1}$ e cun conxunto de valores atípicos.

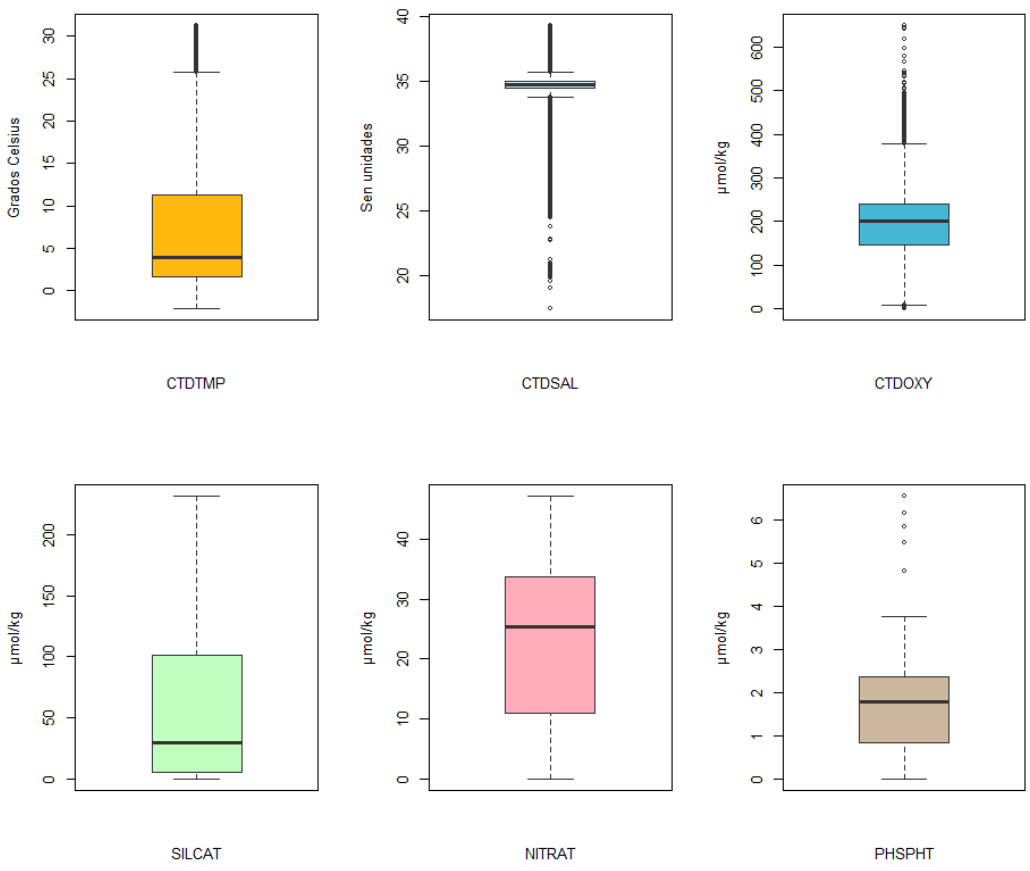

Figura 3.3: Gráfico box-plot das variables físicoquimicas da base de datos de Hansell et al. (2021).

O seguinte gráfico presentado nesta análise exploratoria é un diagrama de barras do número de observacións en cada un dos meses do ano representado na Figura 3.4. Nela obsérvase como existen meses cunha elevada cantidade de mostras, salientando o mes de abril con case 14.000 observacións, e outros meses cun número moito menor. Estes meses con menores observacións acumúlanse, principalmente, nos últimos meses do ano: setembro, outubro, novembro e decembro; con valores que non superan as 4.000 observacións, sendo novembro o mes con menor cantidade. Tamén é destacable o mes de xuño, situado entre diversos meses que conteñen bastantes observacións, pero que conta con algo menos de 4.000.

Como última aproximación analítica das variables deste traballo, móstranse dous gráficos na Figura 3.5 para desvelar a distribución da variable `DEPTH`. Por un lado, temos o gráfico da esquerda que se corresponde cun gráfico *box-plot* onde se observa a presenza de valores atípicos próximos á superficie, e algúns a elevadas profundidades, cunha soa observación por encima dos 8.500 metros. Ademais, a mediana da profundidade sitúase xusto por enriba dos 4.000 metros.

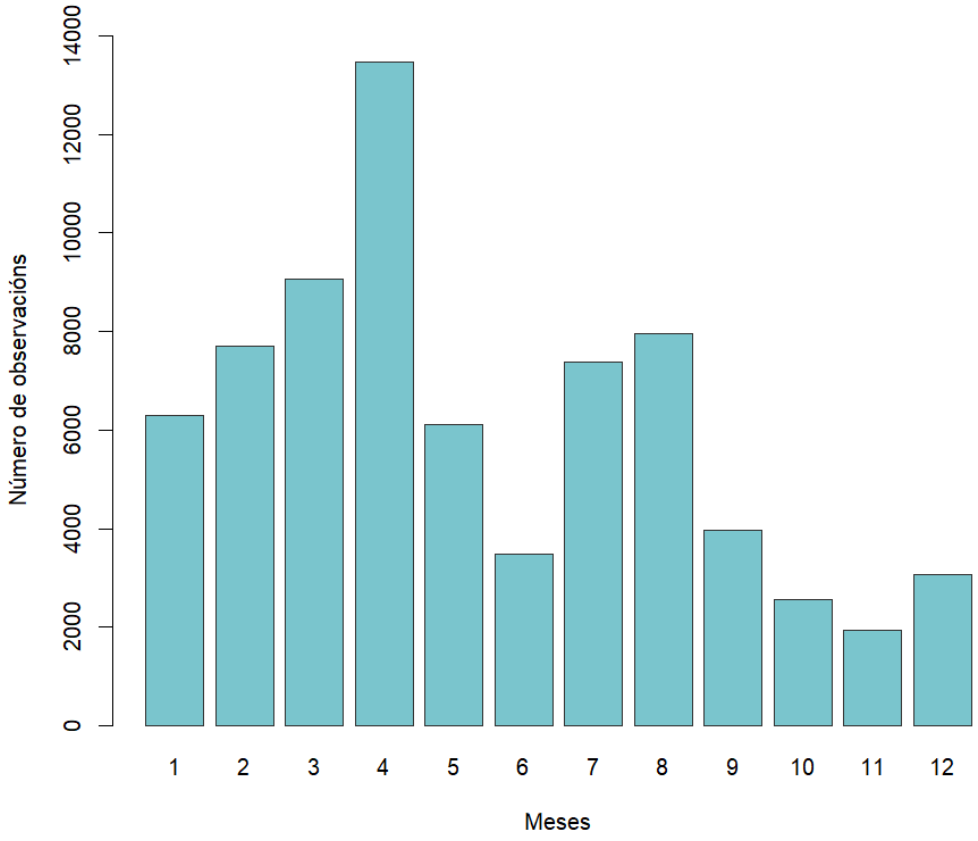

Figura 3.4: Diagrama de barras do nº de observacións da variable mes da base de datos de Hansell et al. (2021).

Ao seu carón, para rematar, obsérvase un histograma onde aparecen reflectidos o número de observacións da variable profundidade. Cada un dos intervalos desta variable correspóndese con 500 metros de amplitude, dividindo a mostra total en 19 intervalos. Pódese observar como o intervalo 5.000 - 5.500 contén máis de 14.000 observacións, sendo o rango con maior cantidade. Este intervalo agrúpase dentro

do rango entre os 3.000 e 6.000 metros de profundidade, que contén a maioría das mostras da base de datos. Pola contra, por encima dos 6.000 metros de profundidade encóntranse moi poucas observacións en comparación cos intervalos previos. Por último, mencionar que no fragmento entre os 500 metros e os 3.000 metros, cada un dos intervalos do histograma ten unha cantidade de observacións similar; mentres que na parte superficial do océano, con profundidades menores que 500 metros e próximas á superficie, rexístranse arredor de 3000 observacións.

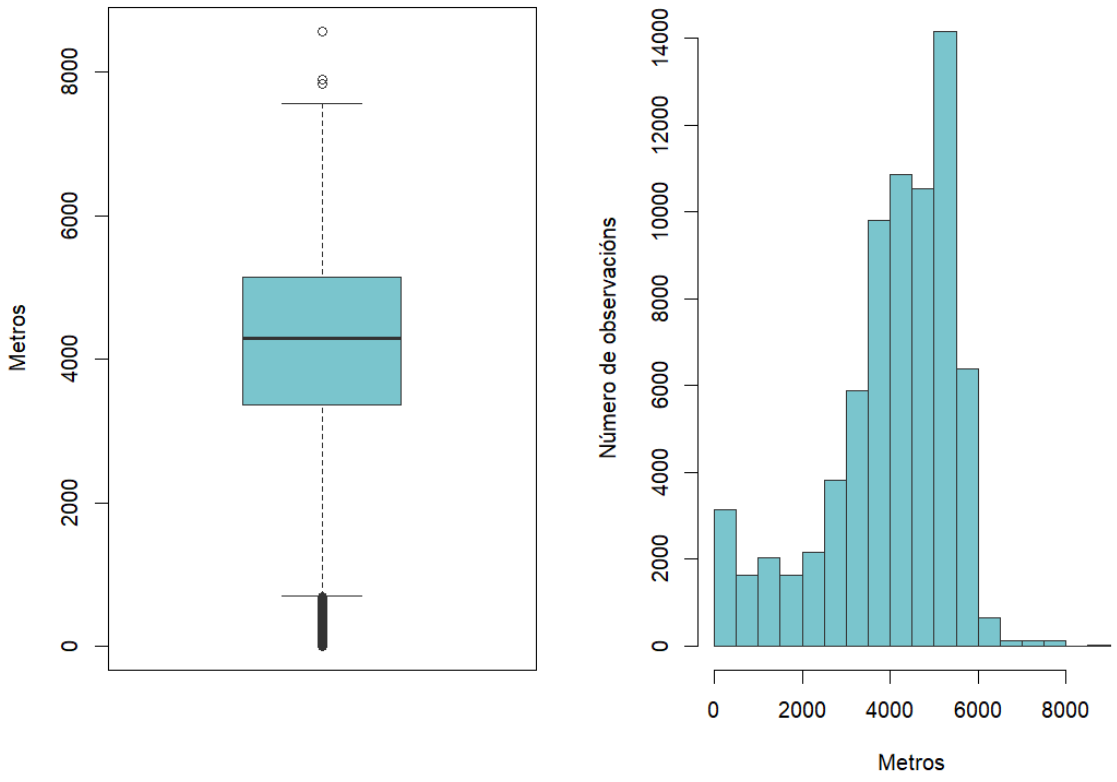

Figura 3.5: Box-plot e histograma da variable `DEPTH` da base de datos de Hansell et al. (2021).

Capítulo 4

Metodoloxía

Neste capítulo do traballo abórdanse cuestións relacionadas cos aspectos máis metodolóxicos. Na primeira sección, descríbense os distintos modelos estatísticos empregados para obter as predicións da variable de interese, o DOC. A sección seguinte céntrase na súa implementación na linguaxe de programación R (R Core Team, 2025) e, finalmente, exponse a validación comparativa entre os diferentes modelos.

Durante todo este capítulo utilizarase como referencia o libro de Hastie et al. (2017) xunto con algúns artigos específicos de cada modelo empregado.

4.1 Modelos estatísticos

Na presente sección descríbese o fundamento teórico e o proceso de axuste dos principais modelos estatísticos empregados neste traballo. En particular, preséntanse catro enfoques distintos: o Modelo Aditivo Xeneralizado (GAM), o método de *Random Forest*, o algoritmo *Extreme Gradient Boosting* e o método dos k-veciños máis próximos (k-NN).

4.1.1 Modelo Aditivo Xeneralizado

De maneira xeral, os modelos de regresión son modelos estatísticos que pretenden representar e explicar a dependencia dunha variable Y, denominada resposta, respecto dunha ou de varias variables $X : \{X_1, X_2, ..., X_p\}$, denominadas explicativas. Ademais de describir esta dependencia entre variables, os modelos constrúense para realizar predicións sobre os valores de Y cando existen novas observacións de X. O método atopa a recta que mellor se axusta, minimizando a diferenza entre os valores preditos e os valores reais da variable resposta. Este modelo multivariante defínese como

$$Y = m(X) + \varepsilon,$$

onde $m(\cdot)$ representa a función media e ε é o erro de regresión.

O modelo de regresión lineal é unha ferramenta estatística amplamente utilizada pola súa sinxela implementación. Este modelo asume que a variable resposta segue unha distribución normal e que existe unha relación lineal entre a variable dependente (resposta) e as covariables. A función de regresión media defínese, neste caso, da seguinte maneira:

$$m(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j,$$

onde $\beta_0, \ldots, \beta_p$ son os parámetros ou coeficientes descoñecidos da regresión.

En moitas situacións, é común que a variable resposta non siga unha distribución normal; neses casos, o uso dun modelo de regresión lineal non é adecuado. No seu lugar, precísase dun enfoque alternativo para modelar a esperanza condicional da resposta. Os Modelos Lineais Xeneralizados (*Generalized Linear Models*, GLM) (Nelder e Wedderburn, 1972), posteriormente discutidos con maior detalle por McCullough e Nelder (1989), ofrecen unha solución ao problema de incorporar diferentes familias de distribucións na variable resposta. Ao estender o modelo lineal, os GLM establecen unha relación entre as covariables e o valor medio da distribución da seguinte maneira:

$$E(Y|X) = m(X) = h^{-1}(\beta_0 + \sum_{j=1}^{p} X_j\beta_j),$$

onde $h(\cdot)$ é unha función monótona coñecida (a función *link*).

A pesar de que os Modelos Lineais Xeneralizados permiten configuracións máis flexibles que o modelo de regresión lineal clásico, a suposición de linealidade entre as covariables e a variable dependente segue sendo moi restritiva. Na actualidade, os modelos non paramétricos son preferidos fronte aos modelos paramétricos e lineais convencionais para modelar a relación entre a resposta e o conxunto de covariables.

Os Modelos Aditivos Xeneralizados (*Generalized Additive Models*, GAM) (Hastie e Tibshirani, 1990) son unha extensión dos GLM que empregan un preditor aditivo composto por funcións suaves e descoñecidas, no lugar do preditor lineal $\eta = \alpha_0 + \sum_{j=1}^{p} \alpha_j X_j$. O preditor aditivo defínese como $\eta = \alpha_0 + \sum_{j=1}^{p} f_j(X_j)$, dando lugar ao seguinte modelo:

$$E(Y|X) = m(X) = h^{-1}(\alpha + \sum_{j=1}^{p} f_j(X_j)), \tag{4.1}$$

onde $h(\cdot)$, coma no caso anterior, é unha función monótona e coñecida (a función *link*), e $f_1, \ldots, f_p$ son as funcións suaves e descoñecidas.

É importante mencionar tamén as condicións de identificación do modelo. Este término refírese á capacidade de determinar de forma única os parámetros do modelo a partir dos datos observados. Un modelo identificable garante unha correspondencia biunívoca entre os parámetros verdadeiros e os datos observados, sendo un factor crucial para asegurar unha interpretación significativa dos parámetros do modelo e realizar inferencias de maneira fiable.

A flexibilidade que ofrecen os GAM na modelaxe de relacións complexas mediante funcións suaves e flexibles pode dar lugar a problemas de non identificación, nos que as diferentes combinacións de parámetros do modelo poden producir o mesmo axuste, o que impide determinar de forma única os verdadeiros valores dos parámetros. Para garantir a identificación do modelo, introdúcese unha constante denotada por α no modelo. Ademais, as funcións parciais deben cumprir a condición $\mathbb{E}[f_j(X_j)] = 0$ para $j = 1, \ldots, p$. Isto implica que $\mathbb{E}[Y] = \alpha$, unha condición crucial para asegurar que as predicións do modelo sexan invariantes incluso ao engadir unha constante a f_1 e ao restar esa mesma constante de f_2 (Hastie e Tibshirani, 1990).

Ademais, os modelos GAM teñen a vantaxe de ser completamente non paramétricos, pero tamén permiten axustar de forma lineal algunhas covariables (por exemplo, para introducir variables categóricas no modelo) dando lugar a un entorno semiparamétrico.

Axuste

A estimación destes modelos pode realizarse empregando distintas funcións suaves non paramétricas, incluíndo *splines* (De Boor, 1978), funcións tipo *kernel* (Wand e Jones, 1994; Fan et al., 1996) ou enfoques bayesianos (Lang e Brezger, 2004).

No caso de utilizar *splines*, pódense representar as funcións suaves de cada X_j en función dunha base lineal coñecida

$$f_j(X_j) = \sum_{i=1}^{n} b_{ij}(X_j)\beta_{ij},$$

sendo $b_i(X_j)$ a i-ésima función da base e β_{ij} un parámetro descoñecido correspondente á j-ésima variable explicativa (X_j) (Liu, 2008).

Para evitar o sobreaxuste e controlar así o grao de suavidade do modelo axustado, engádese unha penalización adicional sobre o axuste a través do método de mínimos cadrados residuais. Porén, dada unha mostra aleatoria $\{(x_i, y_i)\}_{i=1}^{N}$ das variables X e Y, a fórmula que debemos minimizar para axustar o modelo é

$$PRSS(f_j) = \sum_{i=1}^{N}(y_i - f_j(x_i))^2 + \lambda \int (f_j''(x))^2 dx,$$

onde λ convértese nun valor fixo constante. Este parámetro controla o equilibrio entre o axuste do modelo e a suavidade do mesmo. Para estimar o valor óptimo suavizado deste parámetro emprégase a fórmula de validación cruzada xeneralizada (GCV)

$$GCV = \frac{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{f}_\lambda(x_i))^2}{(1 - \frac{tr(S_\lambda)}{N})^2}, \tag{4.2}$$

sendo $tr(S_\lambda)$ a traza da matriz de suavizado do parámetro λ (Liu, 2008).

O algoritmo xeral *Backfitting* (Breiman e Friedman, 1985) presentado no Algoritmo 1 permite axustar un modelo aditivo empregando calquera tipo de función suavizada cando a resposta é Gaussiana. No caso de traballar con variables resposta de outras familias (Binomial, Poisson, etc.) é necesario combinar o *Backfitting* co algoritmo *Local Scoring* (Hastie e Tibshirani, 1990).

Con base no DOC, a variable de estudio deste traballo, a explicación metodolóxica centrarase unicamente no algoritmo *Backfitting*. Este proceso iterativo permite atopar a solución óptima para o modelo proposto (Xiang, 2001). Neste traballo emprégase o modelo da Ecuación en (4.1), sendo a función *link* a identidade, e aplicando splines cúbicos de suavizado que minimizan o método de mínimos cadrados penalizados da Ecuación en (4.2).

Algoritmo 1 Algoritmo de *Backfitting* para modelos aditivos

1: Inicio: $\hat{\alpha} = \frac{1}{N}\sum_{i=1}^{N} y_i$, $\hat{f}_j \equiv 0$, $\forall i, j$.

2: Ciclo: $j = \{1, 2\}, 3, \ldots, p, \ldots, \{1, 2\}, 3, \ldots, p, \ldots$

$$\hat{f}_j \leftarrow S_j\left[\left\{y_i - \hat{\alpha} - \sum_{k\neq j}\hat{f}_k(x_{ik})\right\}_{i=1}^{N}\right]$$

$$\hat{f}_j \leftarrow \hat{f}_j - \frac{1}{N}\sum_{i=1}^{N}\hat{f}_j(x_{ij})$$

ata que a función estimada $\hat{f}_j$ se estabilice.

No paso 1 (Algoritmo 1) fíxase un $\hat{\alpha}$ que nunca cambia. A continuación, aplícase un *spline* suave penalizado S_j sobre $\{y_i - \hat{\alpha} - \sum_{k\neq j}\hat{f}_k(x_{ik})\}_1^N$, en función de x_{ij}, para obter unha nova estimación de $\hat{f}_j$. Este proceso realízase para cada preditor, empregando a estimación das outras funcións $\hat{f}_k$ cando se computa $y_i - \hat{\alpha} - \sum_{k\neq j}\hat{f}_k(x_{ik})$. O proceso continúa ata que a estimación de $\hat{f}_j$ se estabiliza.

4.1.2 Random Forest

As árbores de decisión son un método sinxelo e interpretable de obter as predicións tanto en problemas de clasificación como de regresión (Breiman et al., 1984). Consiste en dividir mediante unha regra de decisión binaria e recursiva todo o espazo do preditor, que é o conxunto de posibles valores correspondentes ás variables explicativas, axustando en cada un deles un modelo simple (Hastie et al., 2017).

As rexións nas que dividimos o espazo do preditor denótanse por R_j, con $j = 1, ..., J$. A idea subxacente é modelar as sucesivas divisións deste espazo creando unha árbore na que cada nodo representa unha división, e cada rama representa as rexións creadas en función desa división ou decisión.

Nun inicio divídense as observacións a través dunha bifurcación que separa o camiño en dúas ramas. Estes puntos de decisión, en función do valor dun preditor, coñécense como nodos internos. Cando o proceso recursivo continúa e chega a un nodo terminal $R_1, ..., R_J$, obtense un valor da variable resposta que consiste no valor medio das observacións nesa rexión R_j.

Durante este proceso é necesario aplicar un método de axuste que nos permita identificar o criterio para dividir o espazo do preditor nas R_j rexións mencionadas. Para iso, dada unha mostra $\{(x_i, y_i)\}_{i=1}^N$, búscase minimizar a suma de cadrados residual (RSS) definida como

$$RSS = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2,$$

sendo $\hat{y}_{R_j}$ a media da variable resposta das mostras de adestramento contidas na rexión R_j.

A pesar de empregar un criterio que optimice as divisións do espazo do preditor, as árbores de regresión poden ser moi sensibles aos datos de entrada, dando lugar a grandes variacións na construción das mesmas. Cun reducido nesgo e unha elevada varianza, parece lóxico aplicar algunha técnica de ensamblado que acople os distintos modelos e permita reducir a varianza do axuste.

Dentro das distintas técnicas de ensamblado, a técnica de *bagging* ou *bootstrap aggragation*, proposta por Breiman (1996), trata de promediar as predicións dentro dunha colección de mostras *bootstrap*, reducindo a súa varianza. Supoñamos que axustamos un modelo coa nosa mostra de adestramento $Z = (x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$, obtendo $\hat{m}(x)$ nun punto x. Para cada unha das mostras *bootstrap* $Z^{*b}, b = 1, 2, ..., B$, axústase a correspondente árbore de decisión e obtense unha predición $\hat{m}^{*b}(x)$. Porén, a estimación *bagging* defínese como

$$\hat{m}_{bag} = \frac{1}{B} \sum_{b=1}^{B} \hat{m}^{*b}(x),$$

onde cada un dos puntos dos nosos datos (x_i, y_i) teñen igualdade de probabilidades: $\frac{1}{N}$.

A idea esencial detrás do *bagging* é promediar a cantidade de ruído pero aproximando modelos innesgados, e polo tanto, reducindo a varianza. Ademais, como cada unha das árbores xeradas no *bagging* atópase identicamente distribuída, o promedio esperado de B árbores é o mesmo que o esperado dun deles soamente. Isto significa que o nesgo das árbores mediante *bagging* é o mesmo que o nesgo das árbores individuais mediante *bootstrap*, entón a única maneira de melloralo é reducindo a varianza. Esta cuestión vai en contra de outra técnica de ensamblado como é o *boosting*, onde os árbores crecen dunha maneira adaptativa para eliminar o nesgo e, polo tanto, non están identicamente distribuídos.

O promedio das B independentes e identicamente distribuídas variables aleatorias, con varianza σ^2, ten varianza $\frac{1}{B}\sigma^2$. Se as variables son simples, identicamente distribuídas pero non necesariamente independentes con unha correlación positiva entre pares ρ, o promedio da varianza é

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2. \tag{4.3}$$

Cando o término B da Ecuación en (4.3) se incrementa, o segundo término desaparece; mais o primeiro mantense e, polo tanto, o tamaño da correlación entre os pares dos árbores calculados mediante *bagging* limita os beneficios de promedialos. Debido a esta cuestión, o método de ensamblado do *bagging* reduce a varianza do modelo final, sempre e cando, os distintos modelos acoplados non estean altamente correlados.

Para asegurar esta restrición, Breiman (2001) propuxo unha modificación substancial da técnica *bagging* denominada *Random Forest*. Esta implementación consiste en mellorar a redución da varianza do *bagging* reducindo a correlación entre as árbores, sen incrementar demasiado a varianza. Isto conséguese no proceso de crecemento das árbores, onde se seleccionan, en cada nodo de decisión, as variables explicativas de maneira aleatoria. O azar na selección de variables preditoras permite que aquelas con maior peso na resposta non dominen, e teñan presenza en todas as árbores xeradas. Por último, débese ter en conta que o número de preditores seleccionados aleatoriamente debe ser inferior ao número total de variables preditoras ($s < p$), senón estariamos aplicando a técnica *bagging* sen ningunha modificación ($s = p$).

Axuste

O Algoritmo 2 permite axustar o modelo *Random Forest* para os problemas de regresión. Para cada unha das submostras Z^* axústase unha árbore de decisión onde, en cada nodo interno, selecciónanse s variables preditoras de xeito aleatorio e, minimizando a suma de cadrados residual (RSS) da Ecuación en (4.1.2), divídese o espazo preditor en dúas rexións. Cando este proceso remata, xerando B árbores de decisión, o resultado da nova observación resulta do promedio destas B árbores xeradas, $T_b(x)$, para a variable explicativa de interese (Breiman, 2001).

Algoritmo 2 Algoritmo para axustar o modelo *Random Forest* para regresión.

1: Dende $b = 1$ ata B:

   (a) Escoller unha submostra Z^* de tamaño N dende os datos de adestramento.

   (b) Axustar unha árbore de decisión T_b sobre a submostra, repetindo de maneira recursiva os seguinte pasos para cada un dos nodos terminais da árbore, ata que se alcanza o mínimo tamaño de nodo n_{min}.

      i. Seleccionar s variables de maneira aleatorias das p variables.

      ii. Escoller a mellor variable ou punto de corte de todas as s.

      iii. Dividir o nodo en dous nodos separados.

2: Obténse o conxunto de árbores $\{T_b\}_1^B$.

3: Predición dunha nova observación mediante o estimador:

$$\hat{m}_{rf}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x).$$

Neste caso, despois de formar o conxunto $\{T(x;\Theta_b)\}_1^B$, Θ_b caracteriza á bth árbore aleatoria en termos da división de variables, puntos de corte en cada nodo e valores do nodo terminal. De maneira intuitiva, ao reducir o valor de s (número de variables seleccionadas en cada submostra), redúcese a correlación entre calquera par de árbores no conxunto e, polo tanto, seguindo a Ecuación en 4.3, redúcese a varianza do modelo final acoplado.

4.1.3 Extreme Gradient Boosting

Os métodos de ensamblado tipo *boosting* son unhas das ideas de aprendizaxe máis potentes desenvoltas nos últimos vinte anos. Propostos no seu inicio como métodos de clasificación, foron máis tarde ampliados cara a regresión. De xeito contrario ao método *bagging*, os modelos axustados iterativamente nesta unión non son independentes nin se poden axustar de forma paralela, dado que teñen en conta aqueles axustados previamente.

Este procedemento traballa con modelos iniciais cun elevado nesgo e unha reducida varianza, e vai avanzando en complexidade incluíndo novas regras de regresión que predín o conxunto de observacións, priorizando aquelas máis nesgadas da regra anterior e reducindo o propio nesgo en cada iteración.

Os métodos de ensamblado permiten empregar distintos modelos de regresión pero, neste traballo, o enfoque céntrase especificamente nas árbores de decisión descritas previamente. Unha das primeiras versións de *boosting*, que tivo bastante éxito nos problemas de clasificación, foi o algoritmo *AdaBoost* (Freund e Schapire, 1995). Anos despois, Friedman (2001) propuxo un novo método denominado *Gradient Boosting*.

O método de ensamblado do *Gradient Boosting* necesita do axuste de árbores de regresión. Así, para representar unha árbore de regresión pode empregarse a seguinte expresión:

$$m(x) = \sum_{j=1}^{J} \gamma_j I(x \in R_j),$$

onde x representa o conxunto de variables explicativas, γ_j refírese á predición que realiza a árbore para as observacións contidas na rexión R_j e I define a función indicadora que toma o valor 1 se $x \in R_j$ e 0, no caso contrario. Dada unha rexión R_j, a estimación de γ_j é: $\hat{\gamma}_j = \bar{y}_j$, a media das observacións y_i desta rexión R_j.

Cada árbore de regresión axustada $\hat{m}(x)$ depende dos parámetros $\Theta = \{R_j, \gamma_j\}_1^J$, sendo J o número de rexións nas que se divide o espazo preditor. Neste método *boosting* búscase atopar unha combinación de parámetros que minimice a seguinte expresión en cada unha das etapas ou árbores:

$$\hat{\Theta} = arg\ min_{\Theta} \sum_{i=1}^{N} L(y_i, m_{t-1}(x_i) + m(x_i)), \tag{4.4}$$

onde $m_{t-1}(x_i)$ representa as predicións da árbore axustada na etapa anterior e $L(\cdot)$ é unha función de perda completamente especificada. Para o contexto da regresión, pode empregarse a suma de cadrados residuais (RSS) definida na Ecuación en (4.1.2).

Por último, a función final obtida a partir do modelo de árbores ensamblado por *boosting* convírtese na suma das árbores dende $t = 1, \ldots, T$ que axusta o algoritmo.

No método de acople *Gradient Boosting* proposto por Breiman (2001), a forma de minimizar a Ecuación en (4.4) é axustar a árbore da etapa m_t en función dos residuos da árbore m_{t-1}, empregando a función

$$r_{it} = -\left[\frac{\partial L(y_i, m(x_i))}{\partial m(x_i)}\right]_{m(x_i)=m_{t-1}(x_i).} \tag{4.5}$$

Ademais, utilizando a medida do erro cadrado para calcular a proximidade das predicións, temos a estimación dos parámetros Θ_m expresados como

$$\tilde{\Theta}_t = arg\ \ min_{\Theta_t} \sum_{i=1}^{N} (-r_{it} - m(x_i))^2,$$

sendo, para esta función de perda do erro cadrado, o gradiente negativo da Ecuación en (4.5) igual aos residuos ordinarios: $-r_{im} = y_i - m_{t-1}(x_i)$.

Recentemente, Chen and Guestrin (2016) desenrolaron un método que se basea no *Gradient Boosting* descrito ata agora, pero que resulta moito máis eficiente, denominado *Extreme Gradient Boosting* (XGBoost). Neste novo acople, estímanse os parámetros definidos na Ecuación en (4.4) empregando os residuos das árbores previas expresados na Ecuación en (4.5). Esta mellora de rendemento conséguese incluíndo en cada iteración a seguinte función penalizada

$$\Omega(m(x)) = \alpha M + \frac{1}{2}\nu\|\omega\|^2,$$

onde M representa o tamaño da árbore, ω o vector das predicións realizadas para cada unha das follas, o valor α controla o crecemento da árbore de xeito inverso e ν é un parámetro de regularización adicional, análogo á taxa de aprendizaxe do *Gradient Boosting*. Deste xeito, incluíndo esta penalización, a función que se debe minimizar para obter os parámetros óptimos en cada iteración é

$$\hat{\Theta} = arg\ min_{\Theta} \sum_{i=1}^{N} L(y_i, m_{t-1}(x_i) + m(x_i)) + \Omega(m(x)).$$

Axuste

O proceso de axuste iterativo do *Extreme Gradient Boosting* aparece reflectido no Algoritmo 3. Na primeira liña se inicia co modelo constante óptimo, que consiste nunha árbore cun só nodo terminal. A continuación, para cada iteración t, dende $t = 1, ..., T$, axústase unha árbore en función dos residuos do modelo anterior e minimizando a función de perda dos valores preditos da variable resposta (γ_{jt}) nesa árbore anterior. Iterativamente, vaise actualizando o modelo ata chegar ao límite T, onde se obtén o resultado final. Por último, sinalar dous hiperparámetros básicos na implantación deste algoritmo, o número de iteracións T e o tamaño de cada unha das árbores axustadas J_t, sendo $t = 1, 2, ..., T$.

4.1.4 k-veciños máis próximos

O algoritmo dos k-veciños máis próximos (*k-Nearest Neighbours*, k-NN) é un algoritmo non paramétrico, sinxelo de entender e que soe ser aplicado para clasificación, mais tamén presenta a súa extensión para o caso da regresión. En ámbolos dous casos, esta técnica emprega a "memoria total" dos datos de adestramento e non realiza ningunha xeneralización sobre eles nin sobre a súa distribución de probabilidades (Song et al., 2017).

O método basease no emprego das observacións máis próximas, dentro do espazo dos datos de adestramento, ao dato x_0 observado para computar a resposta $\hat{Y}$. Igual que outros métodos locais, estas técnicas son coñecidas por funcionar correctamente en bases de datos grandes e con poucas dimensións (Hastie et al., 2017; Kramer, 2013).

Considerando a problemática da regresión, onde temos un conxunto de variables explicativas $X : \{X_1, X_2, ..., X_p\}$ e unha variable resposta Y, o axuste dado mediante os k-veciños máis próximos adquire a forma

$$\hat{Y}(x_0) = \frac{1}{k} \sum_{x_i \in N_k} y_i,$$

onde N_k é a veciñanza de x_0 definida polos k puntos máis próximos x_i na mostra de adestramento. Anotar que o término proximidade implica unha métrica que depende da dimensión dos nosos datos. Neste contexto asumimos distancias euclidianas (en $\mathbb{R}^p$).

Porén, a distancia euclidiana no espazo das características defínese como

$$D_{(i)} = ||(x_0 - x_{(i)}||,$$

Algoritmo 3 Algoritmo de axuste do *Extreme Gradient Boosting.*

1: Inicio co valor $m_0(x) = arg\ min_\gamma \sum_{i=1}^{N} L(y_i, \gamma)$.

2: Para cada $t = 1, .., T$:

(a) Calcular os residuos (gradientes) dados por

$$r_i^{(t)} = -\left[\frac{\partial L(y_i, m(x_i))}{\partial m(x_i)}\right]_{m=m_{t-1}}, \text{para cada } i = 1, ..., N.$$

(b) Axustar unha árbore para eses residuos obtendo as rexións R_{jt} para cada un dos nodos terminais $j = 1, 2, ..., J_t$.

(c) Para cada rexión R_{jt}, calcular

$$\hat{\gamma}_{jt} = arg\ \ min_\gamma \sum_{x_i \in R_{jt}} L(y_i, m_{t-1}(x_i) + \gamma) + \Omega(\gamma),$$

onde $\Omega(f) = \alpha M + \frac{1}{2}\lambda||w||^2$, sendo M o tamaño da árbore, w os valores das predicións nas follas, α un parámetro de penalización que controla o crecemento da árbore e λ un parámetro de regularización adicional.

(d) Actualizar

$$m_t(x) = m_{t-1}(x) + \eta \sum_{j=1}^{J_t} \hat{\gamma}_{jt} I(x \in R_{jt}),$$

onde η é a taxa de aprendizaxe (*learning rate*) e $I(\cdot)$ é a función indicadora.

3: Saída do algoritmo: $\hat{m}(x) = m_T(x)$.

onde x_0 representa o punto ou localización que se quere predicir, $x_{(i)}$ o i-ésimo punto do conxunto de datos observados, e $D_{(i)}$ a distancia euclidiana entre ambos puntos (Kramer, 2013).

A idea dentro da estimación k-NN basease na asunción de que estas funcións locais dentro do espazo dos datos axudan a predicir unha nova observación. Polo tanto, os veciños máis próximos de x_0, restrinxidos por k, deberían conter patróns, valores ou categorías similares que permitan estimar esa $\hat{Y}(x_0)$, sendo modelada segundo o promedio das k observacións máis próximas.

Ademais, no caso da regresión, o tamaño k dos veciños x_i empregados para a estimación da observación x_0 segundo a función $\hat{Y}(x_0)$ ten unha gran relevancia, debido á capacidade de sobreestimar ou infraestimar as estimacións sobre a base de datos no caso de que os valores sexan baixos ($k = 1$) ou elevados ($k = N$), respectivamente (Kramer, 2013).

Axuste

O algoritmo 4 permite modelar a variable resposta Y seleccionando os k veciños máis próximos en cada ciclo de iteración, a partir das observacións da variable explicativa máis próximas no conxunto de datos de adestramento.

Algoritmo 4 Algoritmo de axuste do modelo k-veciños máis próximos

1: Dado un x_0 que se quere predicir (dentro do conxunto inicial da x):

(a) Computar a distancia euclidiana entre o punto dado e as observacións:

$$D_{(i)} = ||(x_0 - x_{(i)}||.$$

(b) Ordenar os valores $D_{(i)}$ de menor a maior.

(c) Seleccionar os k veciños máis próximos a x_0, é dicir, os k puntos con menor $D_{(i)}$ obtendo a submostra N_k.

(c) Estimar a resposta $\hat{Y}(x_0)$ como a media dos valores resposta dos k veciños

$$\hat{Y}(x_0) = \frac{1}{k} \sum_{x_i \in N_k} y_i$$

4.2 Implementación dos modelos

Unha vez expostos os modelos que se empregan neste traballo, procédese a describir a súa implementación no software R xunto coa selección de hiperparámetros realizada no proceso de adestramento dos mesmos. En todos os procesos de busca de hiperparámetros óptimos, e no posterior adestramento, utilízase a base de datos de Hansell (Hansell et al., 2021). Ademais, en todas as implementacións dos diferentes modelos de regresión propostos aplicarase validación cruzada (ver Sección 4.3.1).

4.2.1 Modelo Aditivo Xeneralizado

Para levar a cabo a implementación do modelo GAM emprégase o paquete *mgcv* (Wood, 2017). Esta representa as funcións suaves empregando *splines* de regresión penalizados, utilizando por defecto as funcións base óptimas dado un número delas. Os termos de suavizado poden ser funcións con calquera número de covariantes e, no noso caso, soluciónase o problema da estimación dos parámetros de suavizado empregando o criterio da validación cruzada xeneralizada (GCV). Indicar que as variables `LATITUDE` e `LONXITUDE` inclúense como interacción no modelo mediante o produto tensor suavizado.

4.2.2 Random Forest

Existen diversos paquetes que implementan o modelo *Random Forest* dentro do software R con distintos contextos de aplicación. Neste traballo emprégase o paquete *ranger* (Wright et al., 2024), deseñado como unha rápida implementación dos modelos *Random Forest* e que se atopa aconsellado para traballar especificamente con datos de alta dimensión, reducindo o tempo de computación dos procesos.

Para a selección de hiperparámetros utilízase o paquete *caret* (Kuhn, 2024), debido á ausencia de funcións específicas con este propósito no paquete mencionado anteriormente. A través dun proceso de validación cruzada (*Cross Validation*, CV) procédese á busca da combinación óptima dos seguintes hiperparámetros:

- *mtry*: número de variables posibles nas que dividir cada nodo.
- *min.node.size*: mínimo tamaño dos nodos terminais.

No proceso de axuste e implementación, o hiperparámetro correspondente ao número de árbores foi fixado en $num.trees = 500$. O resto de parámetros mantivéronse cos valores por defecto. A combinación óptima de hiperparámetros obtida é presentada na Táboa 4.1.

Hiperparámetro	**Valor escollido**
num.trees	500
mtry	3
min.node.size	1

Táboa 4.1: Valores escollidos dos hiperparámetros para o axuste e implementación do modelo *Random Forest*.

A continuación, mediante o Algoritmo 2, selecciónanse submostras da base de datos de adestramento para construír cada unha das árbores. Nesta construción, foron seleccionadas s variables de maneira aleatoria en cada un dos nodos. O número de variables escollidas corresponde co hiperparámetro $mtry = 3$ para o noso axuste e, posteriormente, escóllese a mellor variable ou punto de corte óptimo entre todas as s variables escollidas. Isto permite dividir cada nodo en dúas ramas ou nodos separados, ata chegar de maneira recursiva ao mínimo tamaño de nodo seleccionado que corresponde co hiperparámetro $min.node.size = 1$.

4.2.3 Extreme Gradient Boosting

Do mesmo xeito que sucedía anteriormente, para a implementación dos modelos *boosting* en xeral, existen diversos paquetes. Neste caso, para axustar o modelo *Extreme Gradient Boosting* emprégase a interfaz do paquete *caret* para axustar o algoritmo estándar, implementado no método "xgbTree".

Para a busca de hiperparámetros, séguese o mesmo procedemento mencionado para o modelo anterior, empregando o paquete *caret* e a validación cruzada. Neste caso, búscase a combinación máis óptima dos seguintes parámetros:

- *nrounds*: número de iteracións do algoritmo *boosting*.
- *max_ depth*: profundidade máxima da árbore.
- *eta*: parámetro de regularización que controla a tasa de aprendizaxe.
- *colsample_ bytree*: proporción de preditores seleccionados ao azar para crecer cada árbore.
- *subsample*: proporción de observacións seleccionadas ao azar en cada iteración *boosting*.
- *gamma*: mínima redución de perda para facer unha partición adicional nun nodo da árbore.
- *min_ child_ weight*: suma mínima de peso (*hessiana*) para facer unha partición adicional nun nodo da árbore.

O resultado final do proceso de optimización deu como resultado os valores presentados na Táboa 4.2 para cada un dos parámetros.

4.2.4 k-veciños máis próximos

No proceso de implementación deste último algoritmo, empregamos o método "knn" dentro do paquete *caret*. Neste método só é preciso axustar un hiperparámetro:

- k: número de observacións próximas ao valor que queremos predicir.

Hiperparámetro	Valor escollido
$nrounds$	150
max_depth	3
eta	$0,4$
$gamma$	1
$colsample\_bytree$	$0,8$
min_child_weight	$1,8$
$subsample$	1

Táboa 4.2: Valores escollidos dos hiperparámetros para o axuste e implementación do modelo *Gradient Boosting*.

Como sucede nos anteriores procesos de busca, realízase un proceso de validación cruzada para seleccionar o valor óptimo deste parámetro. O resultado obtido aparece representado na Táboa 4.3.

Hiperparámetro	Valor escollido
k	3

Táboa 4.3: Valor escollido do hiperparámetro para o axuste e implementación do *k-NN*.

4.3 Validación dos modelos

O rendemento dun modelo de regresión, e neste caso, dun modelo de aprendizaxe automática, pode relacionarse coa capacidade de predición dun conxunto de datos de proba, novos e independentes. A valoración desta capacidade ou rendemento na predición convértese nun proceso de vital importancia na práctica, debido a que permite escoller o mellor método de aprendizaxe e tamén aporta unha medida da calidade do modelo escollido (Hastie et al., 2017).

Para medir esta xeneralización do método emprégase como medida os erros entre os valores Y e a predición $\hat{m}(X)$ estimada a partir dun conxunto de adestramento τ, caracterizados por unha función de perda denotada como $L(Y, \hat{m}(X))$. Unha vez definida esta función, pódese calcular o erro do modelo na predición dos datos de proba (test), tamén definido como o erro xeneralizado, seguindo a expresión

$$Err_\tau = E[L(Y, \hat{m}(X))|\tau], \tag{4.6}$$

onde X e Y son seleccionados de maneira aleatoria a partir da súa distribución conxunta.

4.3.1 Proceso empregado

Un dos métodos máis simples e máis empregados para estimar os erros de predición é a validación cruzada. Este método estima directamente o erro mediante a fórmula da Ecuación en (4.6) cando se aplica sobre unha mostra test independente. De maneira ideal, se temos suficientes datos, podemos

empregar unha porción dos nosos datos como conxunto test de validación e empregalo para medir o rendemento do noso modelo. Sen embargo, debido a que en numerosas ocasións se dispón dun número limitado de observacións, resulta adecuado empregar a técnica de validación cruzada por bloques, coñecida como *K-fold cross-validation*, para emendar esta dificultade de tamaño na mostra.

Este método utiliza unha parte dos datos dispoñibles para axustar o modelo, e outra parte para validalo. Concretamente, divídese o espazo dos datos en K partes equidistantes (Hastie et al., 2017). Para o fragmento k-ésimo, axústase o modelo coas outras $K-1$ partes restantes do conxunto de datos e calcúlase o erro de predición do modelo cando se predín os valores da resposta do fragmento k-ésimo dos datos. Isto faise para $k = 1, 2, ..., K$ e , finalmente, combínanse as K estimacións para obter o erro de predición. No noso traballo empregamos a situación onde o parámetro $K = 10$.

Dada unha mostra $\{(x_i, y_i)\}_{i=1}^N$ e nomeando como $\hat{m}^{-k(i)}$ a función axustada a partir do conxunto de datos eliminando a fracción k-ésima, a estimación mediante validación cruzada da predición do erro convértese en

$$CV(\hat{m}) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{m}^{-k(i)}(x_i)). \tag{4.7}$$

No proceso de validación cruzada, definido na Ecuación en (4.7), resulta necesario definir unha función de perda que permita cuantificar o erro cometido na predición. Neste traballo considéranse tres formulacións distintas desta función de perda, cada unha das cales dá lugar a un método alternativo para avaliar o rendemento dos modelos. As tres funcións de perda consideradas aparecen na Táboa 4.4.

Nome	Abreviatura	Función de perda $L(Y, \hat{m}(X))$
Mean Absolute Error	MAE	$\frac{1}{N}\sum_{i=1}^{N} \lvert y_i - \hat{y}_i \rvert$
Root Mean Squared Error	RMSE	$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$
R-squared	R^2	$1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \overline{y_i})^2}$

Táboa 4.4: Funcións de perda seleccionadas para cuantificar os erros na estimación das predicións durante o proceso de validación cruzada.

Ademais da estimación dos erros do modelo global, o algoritmo de validación cruzada tamén pode ser empregado na estimación óptima dos hiperparámetros, tal e como foi comentado previamente. Neste caso, dado un conxunto de modelos $m(x, \alpha)$, onde α é o parámetro a optimizar, noméase como $\hat{m}^{-k}(x, \alpha)$ ao α-ésimo modelo axustado coa k-ésima fracción do conxunto de datos eliminada. Porén, para este conxunto de modelos defínese

$$CV(\hat{m}, \alpha) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{m}^{-k(i)}(x_i, \alpha)). \tag{4.8}$$

A función $CV(\hat{m}, \alpha)$ proporciona unha estimación da curva de erro do test, atopando o parámetro óptimo $\hat{\alpha}$ que minimiza esta curva. No proceso de optimización dos hiperparámetros dos modelos empregados neste traballo, utilízase un proceso de validación cruzada onde $K = 5$.

4.3.2 Resultados da validación dos modelos

Na Táboa 4.5 preséntanse os resultados obtidos mediante a aplicación da validación cruzada *K-fold*, empregando $K = 10$ particións para avaliar o rendemento dos modelos previamente descritos. Como se pode observar, o modelo axustado que presenta menores erros e, polo tanto, un mellor axuste, é o modelo *Random Forest*. Tendo en conta tódalas métricas empregadas para avaliar o erro nas predicións (MAE, RMSE, R^2), este modelo é o que acada os mellores resultados en termos de precisión, aínda que non é o modelo máis eficiente en canto ao tempo de computación.

Modelo	Paquete	Tempo (s)*	MAE	RMSE	R-squared
GAM	*mgcv*	**14,08**	2,422	4,195	0,863
Random Forest	*ranger*	70,75	**1,496**	**2,750**	**0,941**
Extreme Gradient Boosting	*caret*	32,89	1,870	3,173	0,921
k-NN	*caret*	303,14	2,045	4,003	0,875

* *Os tempos de computación son aproximados, foron calculados de maneira recursiva sen ningún outro proceso en activo, só co software R.*

Táboa 4.5: Resultado do algoritmo da validación cruzada *K-fold* empregado para a validación dos modelos de regresión.

En situacións opostas temos o modelo GAM e o modelo k-NN: aínda que obteñen resultados semellantes no RMSE e no coeficiente de determinación (R^2), amosan diferenzas notables no tempo de computación. Concretamente, o modelo GAM destaca por ser o máis eficiente neste aspecto, mentres que o algoritmo k-NN, implementado a través da libraría *caret*, require un tempo de execución bastante superior ao do resto dos modelos analizados.

Finalmente, analízase o modelo *Extreme Gradient Boosting*, que presenta un tempo de computación inferior —aproximadamente a metade do precisado polo modelo *Random Forest*—, mantendo ao mesmo tempo uns resultados en termos de RMSE e R^2 relativamente próximos. Polo tanto, en situacións nas que se busque reducir o custo de computación sen comprometer substancialmente a calidade das predicións, este modelo constitúe unha opción especialmente recomendable.

Aínda así, obsérvanse certas diferenzas entre estes dous modelos tanto nos dous parámetros de cuantificación do erro mencionados (RMSE e R^2) como no MAE, onde son máis evidentes as diferenzas e o *Random Forest* mostra uns resultados moito máis precisos.Tendo en conta os resultados obtidos, a elección do modelo de regresión máis axeitado para a estimación da variable DOC resulta clara: o modelo *Random Forest* ofrece o mellor equilibrio entre precisión e rendemento.

Capítulo 5

Resultados

Despois de ter realizado todo o proceso de axuste, implementación e validación dos modelos de regresión, neste capítulo unicamente se presentan os resultados obtidos do axuste do modelo *Random Forest* aos datos de Hansell. Tamén se obteñen as predicións do DOC utilizando a base de datos de WOA, en función das 10 variables explicativas en cada un dos puntos de mostra, nos diferentes meses e ao longo das 40 profundidades. Na representación dos resultados, ademais da explicación e interpretación do modelo de regresión óptimo, engádense certas climatoloxías mensuais en diferentes profundidades e seccións verticais oceánicas das predicións do DOC para exemplificar, comparar e xustificar a capacidade de predición do axuste proposto neste traballo.

5.1 Interpretación do modelo

Como se describiu durante o Capítulo 4, o modelo de aprendizaxe automática *Random Forest* constrúe unha gran cantidade de árbores para logo promedialos e obter as estimacións finais da variable resposta. Con este modelo obtivéronse os mellores resultados, mostrados na Táboa 4.5 da Subsección 4.3.

Na Figura 5.1 aparece exemplificada a primeira árbore aleatoria do modelo *Random Forest* axustado coa base de datos de Hansell (Subsección 4.2). A información representada permite coñecer as divisións realizadas na árbore en base ás covariables consideradas, é dicir, permite coñecer en cada nodo que covariable foi seleccionada xunto coa división do espazo dando lugar a dous novos nodos ou rexións preditoras. Por exemplo, o primeiro nodo (nodo 0) divídese segundo a variable e valor $\texttt{SILCAT} < 7.879$. Deste xeito, vaise continuar cara o nodo 1 ou 2 en función do valor da variable explicativa `SILCAT` nesa nova observación. No caso de ser maior que o valor 7.879, o proceso desprázase cara o nodo 2 ($\texttt{NITRAT} < 15.845$); pola contra, se é menor que este valor, diríxese cara o nodo 1 ($\texttt{CTDTMP} < 20.708$). Este proceso avanza polos diferentes nodos, onde en cada un deles se produce unha división ou corte segundo a variable e o valor óptimos escollidos aleatoriamente nese nodo.

Este sistema de división a través de nodos vai construíndo unha árbore complexa e moi profunda, cunha gran cantidade de variables e nodos ata chegar ao mínimo tamaño de nodo considerado no axuste. Sen embargo, polo camiño tamén vai obtendo valores da variable resposta en función dos datos empregados para adestrar este modelo. Isto aparece representado no nodo terminal desta primeira árbore da Figura 5.1. Estes nodos terminais ofrecen valores de predición da variable resposta do modelo, empregando os valores das variables explicativas da nova observación que queremos predicir. Neste caso, o valor da predición sería: $\texttt{DOC} = 80.80$.

Por último, como en cada unha destas árbores aleatorias non se seleccionan todas as mostras de adestramento ao realizar submostras *bootstrap* no inicio da construción de cada árbore, finalmente

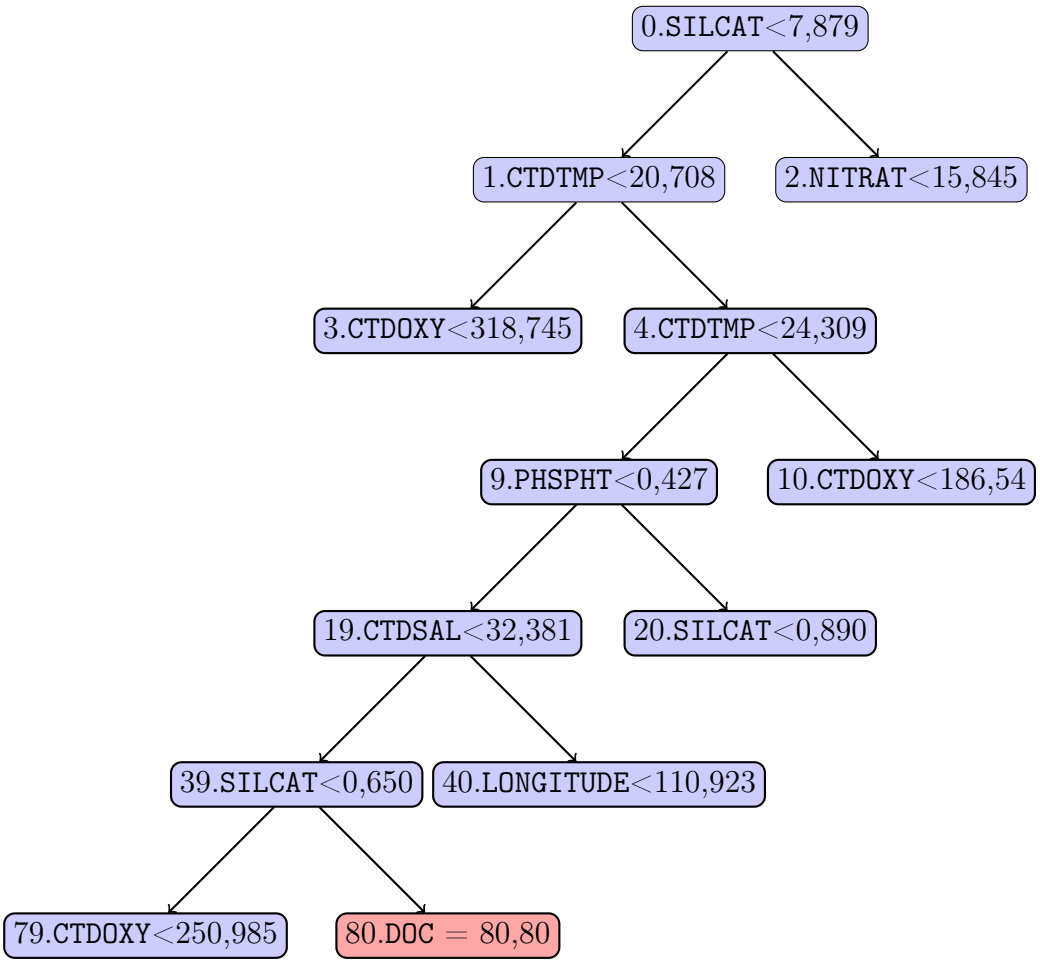

Figura 5.1: Estrutura da primeira árbore aleatoria do modelo *Random Forest* co paquete *ranger*.

realízase unha media dos valores da resposta obtidos en cada unha das 500 árbores aleatorias realizadas. Porén, o valor da estimación da variable resposta en cada unha das novas observacións é un promedio dos valores obtidos en cada unha das árbores aleatorias construídas; garantindo que o azar da selección das explicativas en cada un dos nodos reduce o valor da varianza do modelo e mantén o nesgo de cada unha das árbores individuais.

Como foi mencionado anteriormente, durante o proceso de construción das árbores aleatorias, as variables seleccionadas en cada un dos nodos son escollidas de maneira aleatoria, aínda que posteriormente, entre as s variables, se faga unha selección do mellor punto de corte. A pesar disto, cada unha das variables actúa de forma independente e, polo tanto, pode ter diferente relevancia na predición da variable resposta. Porén, durante o axuste do modelo *Random Forest* calcúlase tamén a importancia de cada unha das variables na predición da variable `DOC`, mostrando o resultado na Figura 5.2. O cálculo da importancia de cada unha das variables realízase a través da varianza das respostas para a variable correspondente (Wright et al., 2024).

Os resultados presentados nesta figura mostran como as variables con maior importancia na predición da variable resposta son o nitrato (`NITRAT`) e, en menor medida, o fosfato (`PHSPHT`). A continuación, aparecen unha serie de variables cunha importancia intermedia como son os silicatos (`SILCAT`), a temperatura (`CTDTMP`) e a salinidade (`CTDSAL`). Destacar que todas as variables mencionadas ata agora

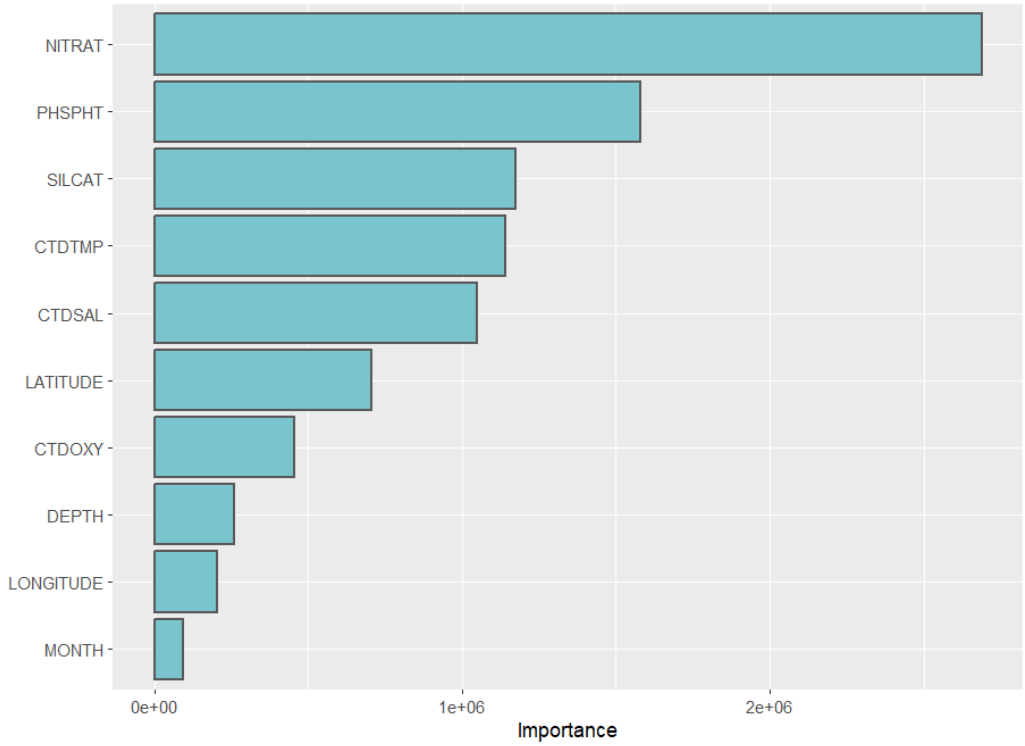

Figura 5.2: Gráfico da importancia relativa das variables na predición da variable `DOC` empregando o modelo *Random Forest* axustado mediante o paquete *ranger*.

son características químicas ou físicas do océano. Na parte final do gráfico, con moi pouca repercusión, aparecen as variables da profundidade (`DEPTH`), a lonxitude (`LONGITUDE`) e o mes (`MONTH`). Por último, nun nivel por encima destas, reflexando algo máis de importancia sobre a resposta, atopamos a latitude (`LATITUDE`) e o osíxeno disolto (`CTDOXY`).

Nesta Figura 5.2 destaca que a variable `NITRAT` ten unha gran importancia á hora de elaborar o modelo de regresión para estimar as concentracións do DOC no océano global. Debido a esta alta dependencia, pode ser relevante observar máis en profundidade a interacción entre os valores desta variable preditora e os valores da variable resposta. Esta relación preséntase no gráfico esquerdo da Figura 5.3. Nela represéntase a función de dependencia parcial, implementada mediante o paquete *pdp* (Greenwell, 2024), da variable `NITRAT` fronte aos valores estimados da variable resposta. Pódese observar como os valores máis altos do `DOC` coinciden cos valores máis baixos da variable `NITRAT` e que a medida que van aumentando os valores da variable `NITRAT`, redúcense os valores da resposta ata chegar a unha estabilización a partir dos 30 $\mu molkg^{-1}$.

Por outra banda, no gráfico da dereita da Figura 5.3 tamén aparece a función de dependencia parcial da segunda variable máis importante: `PHSPHT`, aínda que ten unha menor relevancia. Pódese observar como os valores máis elevados do `DOC` coinciden cos valores máis baixos da variable `PHSPHT`, tal e como acontecía coa variable dos nitratos. Sen embargo, este efecto desaparece cando os valores da `PHSPHT` alcanzan aproximadamente 2 $\mu molkg^{-1}$, onde os valores da variable resposta chegan a un mínimo e comezan a aumentar a medida que aumenta a variable preditora. Este aumento remata cando a variable explicativa chega arredor dos 5 $\mu molkg^{-1}$, onde se estabilizan os valores da variable resposta.

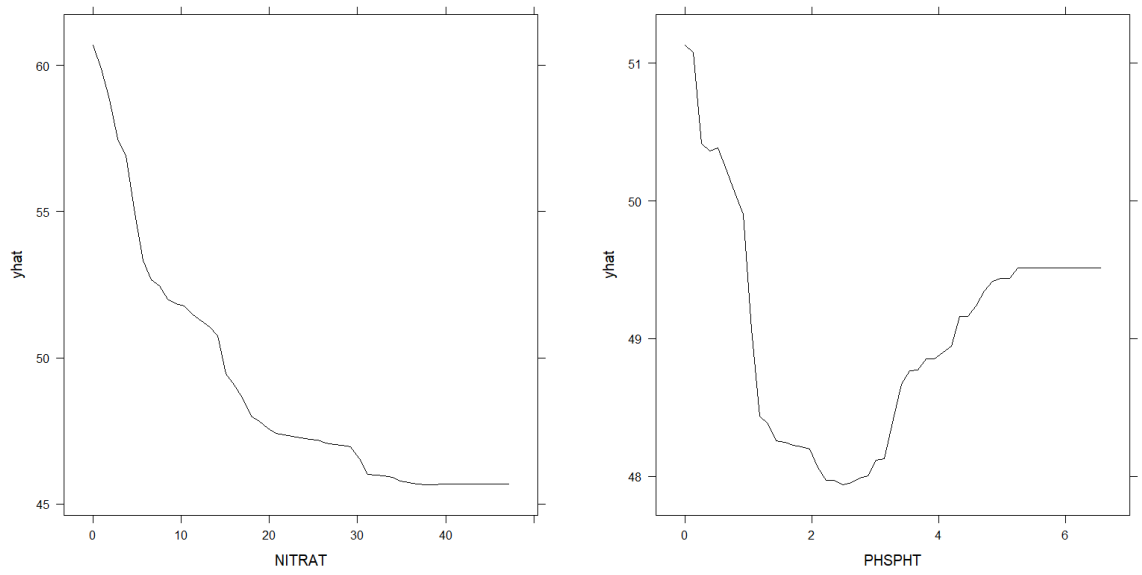

Figura 5.3: Gráfico das funcións de dependencia parcial (*partial dependence plot*, pdp) das dúas variables con maior importancia na predición da resposta segundo o modelo *Random Forest* axustado.

5.2 Predición do DOC

Como xa se mencionou con anterioridade, todo o proceso de axuste do modelo coa base de datos de adestramento de Hansell, realizouse co obxectivo de predicir e visualizar a distribución desta compoñente orgánica no océano global ao longo dos diferentes meses do ano e, sobre todo, a súa variación nas diferentes profundidades do océano.

Porén, neste apartado, vanse presentar algúns exemplos de climatoloxías elaboradas a partir da predición das concentracións do DOC para ser comparadas con outras climatoloxías atopadas na bibliografía, ben sexa observadas *in situ* ou elaboradas a partir doutra metodoloxía. Ademais, tamén se comparan seccións verticais correspondentes a diversas investigacións oceánicas onde as concentracións do DOC proceden de mostraxes e son coñecidas, coa intención de comparar a efectividade e capacidade preditora do noso modelo.

5.2.1 Climatoloxías globais

Empregando a base de datos de WOA e o modelo *Random Forest* adestrado, obtivéronse as climatoloxías para as 40 profundidades e os 12 meses do ano, o que supón un total de 480. Dado o elevado número de resultados, neste apartado preséntanse as climatoloxías globais da predición das concentracións do carbono orgánico disolto (DOC) para determinadas profundidades e algúns meses concretos, co obxectivo de comparar os resultados con outros artigos e referencias bibliográficas, a fin de avaliar a adecuación do noso modelo. Todas as climatoloxías que se presentan foron xeradas mediante o software Panoply netCDF, HDF and GRIB Data Viewer, un software de visualización desenvolto pola NASA *Goddard Institute for Space Studies* (NASA GISS, 2023).

Dos resultados obtidos, pódese resaltar que as predicións do modelo a 0 metros de profundidade non varían de maneira substancial en función do mes no que se produzan. Por exemplo, isto pode observarse na Figura 5.4, Figura 5.5 e Figura 5.6, correspondentes a Marzo, Xullo e Outubro, respectivamente. Obsérvanse diferenzas na distribución das concentracións nas rexións subtropicais e próximas ao ecuador

pola variabilidade estacional, mais de maneira xeral non se modifican as localizacións con maiores e menores concentracións a nivel superficial.

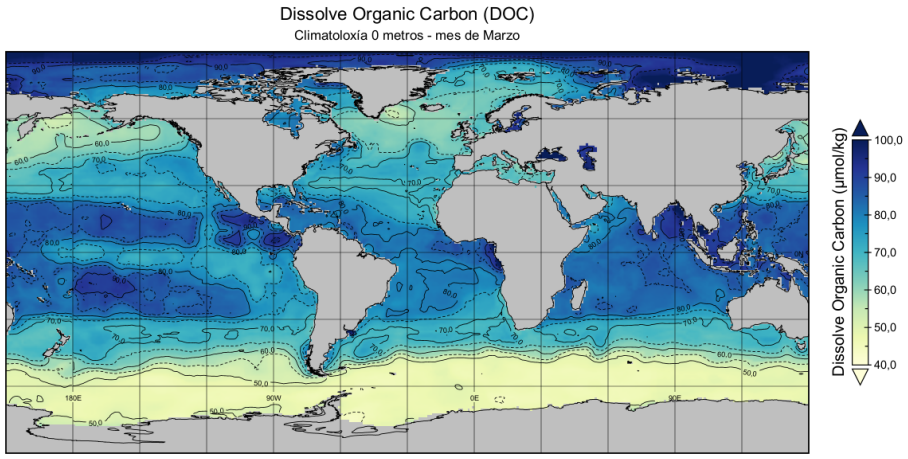

Figura 5.4: Climatoloxía das estimacións do DOC en superficie (0 metros), correspondentes ao mes de Marzo.

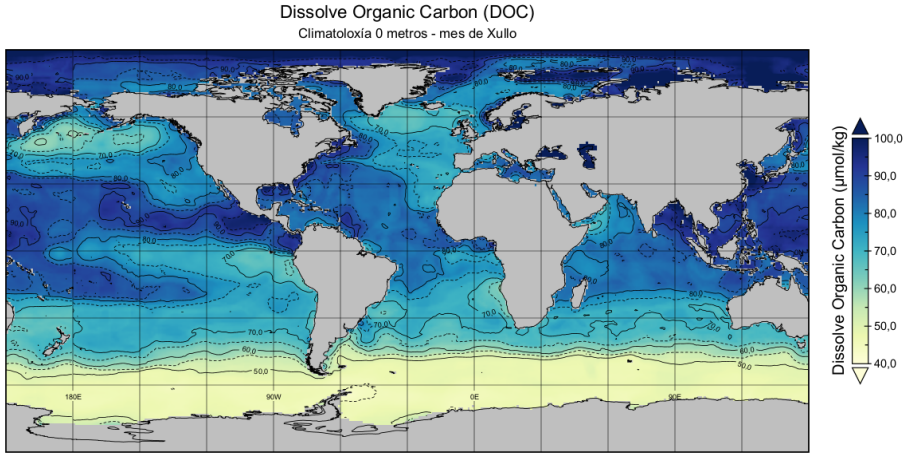

Figura 5.5: Climatoloxía das estimacións do DOC en superficie (0 metros), correspondentes ao mes de Xullo.

A continuación, serán presentadas gráficas onde as predicións do DOC correspondan aos mesmos meses empregados nos artigos de referencia e que serán o obxecto de comparación e validación das predicións. A primeira climatoloxía (Figura 5.6) pode ser comparable ás presentadas no artigo de Laine et al. (2024), onde tamén utilizan o modelo *Random Forest* pero empregando datos satélite á vez que medidas *in situ* para adestrar o modelo. Tamén pode ser enfrontada ás recentes climatoloxías do artigo de Panaïotis et al. (2025), concretamente, á climatoloxía correspondente ao nivel superficial $(0 - 10\text{m})$ do océano; neste último caso empregan o modelo de *Boosted Regression Trees* (BRTs). Os resultados mostran unhas concentracións altas $(70 - 80\ \mu molkg^{-1})$ en océano aberto nas rexións próximas ao ecuador nos diferentes océanos, tanto no Atlántico coma no Pacífico e Índico, xunto con valores moi altos nas rexións próximas á costa e próximas ás desembocaduras de grandes ríos, uns datos acordes aos

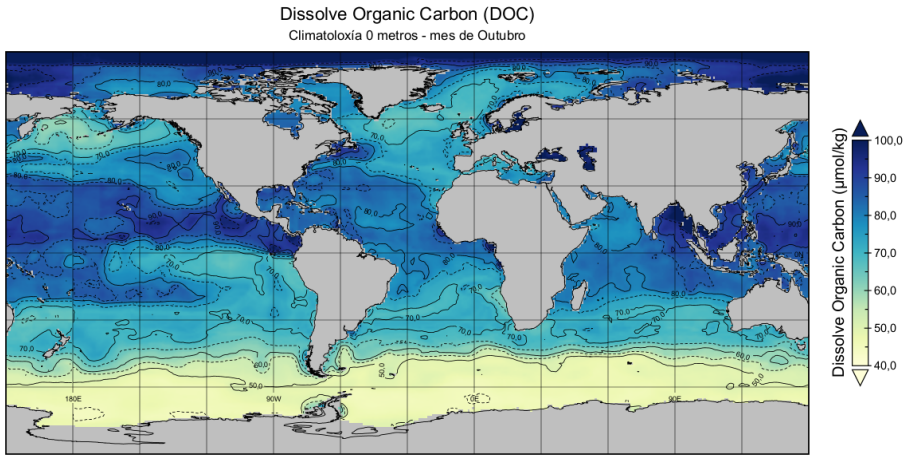

Figura 5.6: Climatoloxía das estimacións do DOC en superficie (0 metros), correspondentes ao mes de Outubro.

resultados das investigacións reflectidos en Hansell et al. (2024) e nos dous artigos mencionados (Laine et al., 2024; Panaïotis et al., 2025). Destácanse tamén, as concentracións superiores aos 90 $\mu molkg^{-1}$ no golfo de Bengala (lonxitude 90°E), no mar Báltico (lonxitude 20°E) ou na rexión do Pacífico próxima á costa oeste de América (lonxitude -100°E) en latitudes próximas ao ecuador. Ademais, os valores máis baixos, arredor dos $40 - 45$ $\mu molkg^{-1}$, observámolos na rexión pertencente ao Antártico (latitude -60°N) e que se corresponden co inicio das correntes AAIW[1] cara as profundidades, arredor dos 1.000 metros, de latitudes intermedias do océano Atlántico.

Tamén se poden destacar as concentracións próximas aos $70 - 75$ $\mu molkg^{-1}$ que atopamos na rexión do Atlántico Norte (latitude entre os 30°N e os 60°N e lonxitude -30°E) xunto cunhas elevadas concentracións en toda a rexión Ártica (latitude por encima dos 70°N) (Hansell et al., 2009), seguindo a lóxica do proposta na introdución do traballo e recollido amplamente pola bibliografía. Por último, mencionar as reducidas concentracións atopadas na rexión norte do Océano Pacífico (lonxitude 180°E e latitude 50°N), próxima ao estreito de Bering, onde se localizan concentracións arredor dos 60 $\mu molkg^{-1}$, bastante diferentes aos 70, 80 e, incluso, 90 $\mu molkg^{-1}$ que predí o modelo nas rexións intermedias deste océano. Todas estas estimacións e patróns oceánicos mencionados ofrecen resultados moi semellantes á climatoloxía superficial descrita por Panaïotis et al. (2025), a pesar de empregar outro modelo distinto na súa construción e unha cantidade moito menor de profundidades.

Na comparativa coas climatoloxías mensuais presentadas por Laine et al. (2024), os resultados obtidos neste traballo outorgan unha maior profundidade e amplitude das climatoloxías, abordando todo o océano global e captando diferenzas máis precisas nas concentracións do DOC entre distintas rexións do océano a nivel superficial.

Pasamos agora a analizar as climatoloxías representadas na Figura 5.7 para o mes de Marzo tendo en conta dúas profundidades: 30 metros (a) e 3000 metros (b). Nesta Figura 5.7, replícanse as dúas climatoloxías incluídas no artigo de Hansell et al. (2009) elaboradas con datos do DOC *in situ* xunto con outras estimacións empregando un modelo bioxeoquímico/físico desenvolto por Schlitzer (2002, 2007). Na parte superior da Figura 5.7 (a) obsérvanse valores moi próximos aos da Figura 5.6, debido á pouca diferenza de profundidade. Áchanse concentracións maiores a 80 $\mu molkg^{-1}$ nas rexións subtropicais dos diferentes océanos e, pola contra, obsérvanse menores concentracións no Atlántico Norte e na zona

[1] AAIW: Antartic Intermediate Water. Correntes que afunden dende a superficie do Océano Austral pouco enriquecidas en DOC.

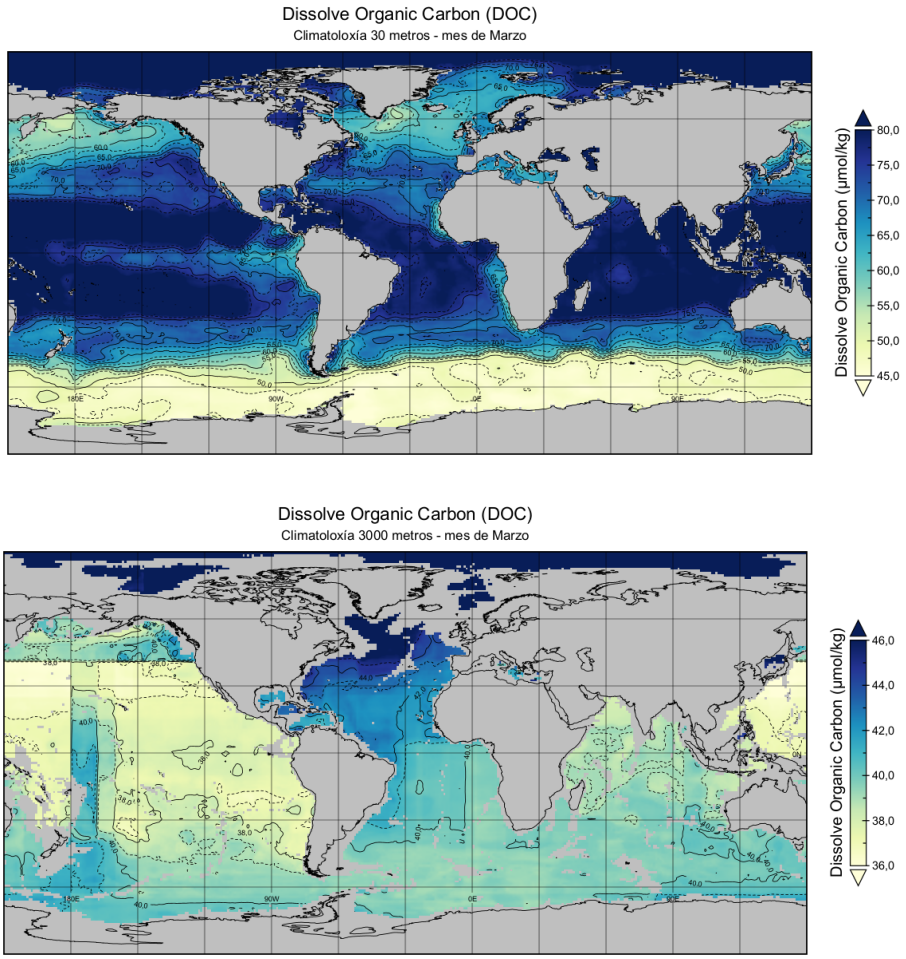

Figura 5.7: Climatoloxías das estimacións do DOC relativas ao mes de Marzo nas profundidades de (a) 30 metros e (b) 3.000 metros.

do Pacífico Norte e en latitudes entre os 30°N e 60°N, arredor dos $55 - 60$ $\mu molkg^{-1}$. Todos estes rexistros aseméllanse á climatoloxía a 30 metros de profundidade presentada por Hansell et al. (2009); quizais mostrando menores concentracións no océano aberto en rexións próximas ao ecuador e nas zonas subtropicais, onde as estimacións neste traballo ofrecen concentracións algo máis elevadas.

Sen embargo, tamén podemos observar certas diferenzas nos valores destas concentracións: destacando as concentracións próximas aos 65 $\mu molkg^{-1}$ atopadas na costa oeste de América (latitude -90°E), arredor do ecuador no Océano Pacífico, que non son propostas no artigo de Hansell et al. (2009). Do mesmo xeito que certas rexións do Atlántico na costa oeste de África (lonxitude -15°E e latitude 20°N), onde o noso modelo predí menores concentracións que no océano aberto. Por último, destacar que na rexión próxima a Australia e ás illas do Pacífico, o modelo estima concentracións máis elevadas que as propostas na investigación de referencia (Hansell et al., 2009).

De xeito paralelo, esta climatoloxía elaborada a 30 metros de profundidade pode servir de comparación para a climatoloxía media da zona epipeláxica (10-200 metros), presentada por Panaïotis et al. (2025). Obsérvanse grandes similitudes entre as dúas climatoloxías, captando correctamente as maiores concentracións do océano aberto nas zonas subtropicais, menores concentracións no Atlántico e Pacífico Norte,

as concentracións máis baixas no Océano Antártico e, por exemplo, tamén predí adecuadamente as concentracións da rexión ecuatorial da costa oeste de América (lonxitude $-90°$E) que se adentran cara ao Pacífico (Figura 5.7).

Por outra banda, na segunda climatoloxía (3000 metros) situada na parte inferior da Figura 5.7 (b) pode observarse que o modelo predí correctamente as concentracións e o seu gradiente ao longo do Atlántico de norte a sur. Debido ao afundimento das masas de auga oceánicas superficiais na rexión do Atlántico Norte durante a corrente circulatoria meridional, no proceso de formación da corrente NADW[2], atopamos concentracións de DOC arredor dos $46-48$ $\mu molkg^{-1}$ como aparece mostrado na Figura 5.7 e tamén recollido no artigo de Hansell et al. (2009). A medida que avanzamos cara o sur, as concentracións redúcense e mantéñense constantes arredor dos 40 $\mu molkg^{-1}$ en toda a rexión do Atlántico Sur e nas zonas do Océano Índico e Océano Pacífico próximas á Antártida. Estas concentracións tamén se ven afectadas pola chegada da corrente NADW que aflora en latitudes próximas á Antártida. Ademais, predí adecuadamente as elevadas concentracións nas rexións Árticas (maiores que 50 $\mu molkg^{-1}$) e tamén aquelas no Océano Pacífico onde se sitúa a corrente profunda PDW[3] cunhas concentracións de 38 $\mu molkg^{-1}$. Este movemento de augas moi pouco enriquecidas mantén toda a rexión Pacífica cunhas concentracións moi reducidas, que o modelo predí de maneira bastante regular.

Pola contra, tamén se atopan dúas rexións equivocadas no Océano Pacífico. A primeira localízase ao redor dos $-180°$E de lonxitude, entre as latitudes -30°N e 30°N; mentres que a segunda abarca toda a franxa correspondente aos 45°N de latitude. En particular, na primeira rexión, o noso modelo estima concentracións excesivamente elevadas de 40 $\mu molkg^{-1}$ quizais pola falta de datos no adestramento para estas profundidades. Deste xeito observamos como, en profundidades arredor dos 3.000 metros, o noso modelo axustado non actúa adecuadamente e elimina a continuidade arredor do globo terráqueo que se podía observar en climatoloxías menos profundas, como a outra da Figura 5.7 ou a Figura 5.6. No caso da segunda rexión, esta coincide cunha elevada densidade de puntos de mostra (ver Figura 3.1) e atópase localizada en latitudes próximas á rexión Ártica, onde no Océano Atlántico se observan altas concentracións de DOC. Ambos factores poden ser cruciais na xeración de predicións pouco axustadas nesta rexión do Océano Pacífico, reflectindo concentracións superiores ás esperadas e afastadas das climatoloxías recollidas en Hansell et al. (2009) para eses niveis de profundidade.

A última figura relativa ás climatoloxías globais correspóndese con tres climatoloxías nas profundidades de (a) 20 metros, (b) 300 metros e (c) 600 metros estimadas para o mes de abril (Figura 5.8). Estas representacións adecúanse ás elaboradas por Roshan e DeVries (2017) onde empregan un modelo baseado en redes neuronais artificiais (ANN) para estimar os valores do DOC só nestas profundidades. O motivo da escolla deste mes débese á maior cantidade de observacións como aparece na Figura 3.4. De maneira xeral, as predicións propostas polo noso modelo presentan unha distribución semellante ás figuras do artigo de Roshan e DeVries (2017) e, polo tanto, ao resto da bibliografía sobre o DOC. Destacando unha maior profundidade e variabilidade na predición destas concentracións, resultando nunha aproximación máis ampla e completa. Nunha visión máis concreta das climatoloxías pódese destacar como a nivel case superficial, nos 20 metros (Figura 5.8 (a)), o modelo capta as maiores concentracións no océano aberto arredor das zonas subtropicais e próximas ao ecuador e as menores concentracións nas rexións do Atlántico Norte e Pacífico Norte, sen esquecer as rexións Ártica e Antártida, onde estima concentracións moi elevadas (>80 $\mu molkg^{-1}$) e constantes sobre os 45 $\mu molkg^{-1}$, respectivamente. A medida que aumentamos a profundidade, na climatoloxía (b) correspondente aos 300 metros, atopámonos con que as maiores concentracións sitúanse na rexión do Atlántico Norte, no punto de formación da corrente profunda do Atlántico Norte (NADW). Ademais desta rexión, obsérvanse concentracións medias sobre os $55-60$ $\mu molkg^{-1}$ nas rexións subtropicais dos diversos océanos, de maneira similar ao que rexistra o artigo de Roshan e DeVries (2017). Nas rexións próximas á Antártida atopamos concentracións constantes e baixas ($40-42$ $\mu molkg^{-1}$) debido á presenza das correntes AAIW cara latitudes máis

[2]NADW: North Atlantic Deep Water. Corrente superficial que afunde no Atlántico Norte e diríxese cara zonas subtropicais, formando parte da corrente termohalina.

[3]PDW: *Pacific Deep Water*. Corrente das profundidades do Pacífico que aflora arredor dos 3.000 metros.

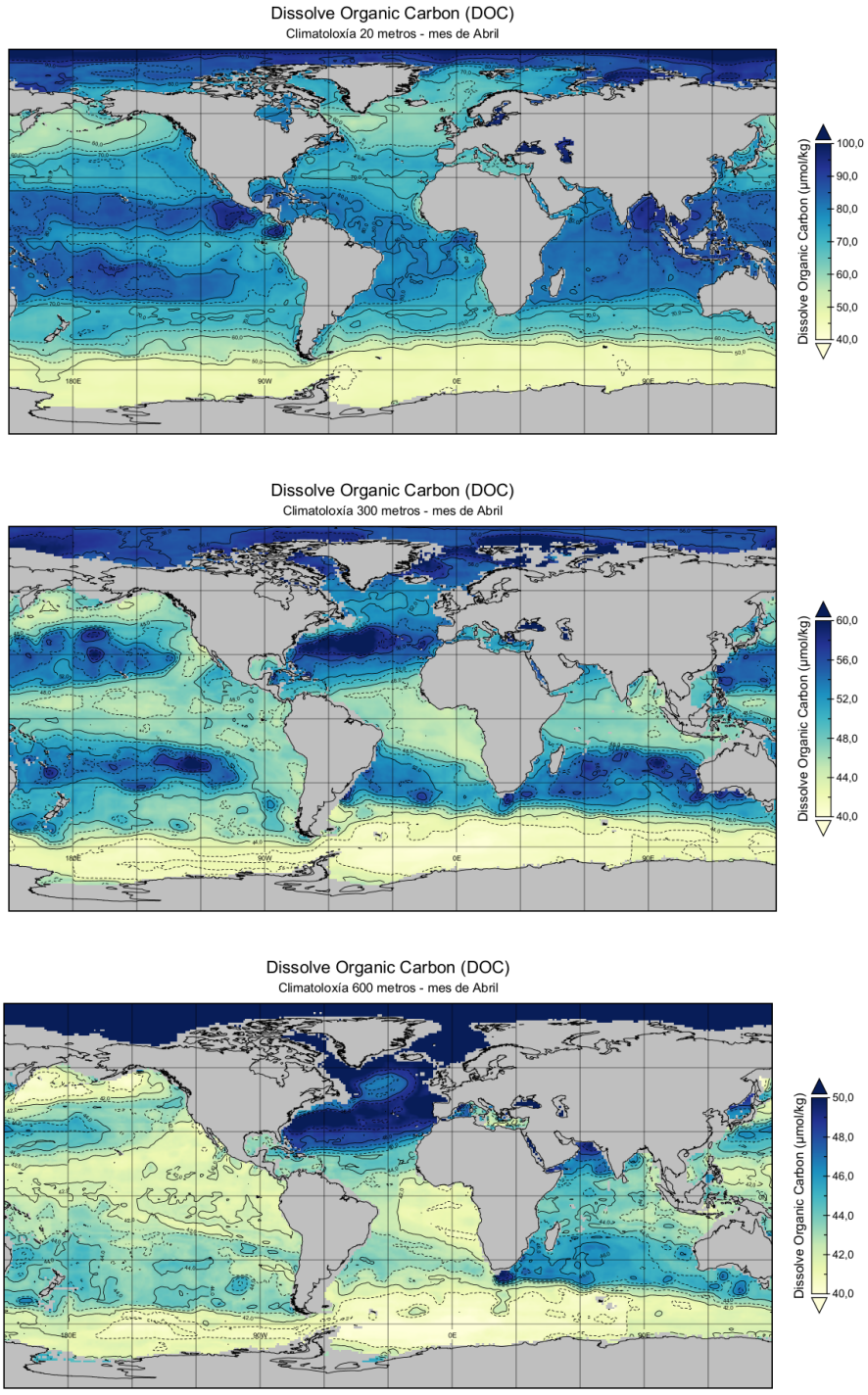

Figura 5.8: Climatoloxías das estimacións do DOC relativas ao mes de Abril en diferentes profundidades a partir do modelo *Random Forest* axustado. As profundidades escollidas correspóndense coas mesmas que o artigo de Roshan e DeVries (2017): (a) 20 metros, (b) 300 metros e (c) 600 metros.

intermedias do Océano.

Por último, na climatoloxía (c) relativa aos 600 metros da Figura 5.8, denotamos con maior claridade o gradiente de concentracións do Océano Atlántico que fomos comentando ata agora. Nestas profundidades as diferenzas non son elevadas en unidades pero si existe moita variación entre as diferentes áreas. Mentres que na zona de formación da NADW no Atlántico Norte e no Ártico se atopan concentracións por enriba dos 50 $\mu molkg^{-1}$, no resto do Océano en latitudes próximas e por debaixo do Ecuador as estimacións son de $44 - 42$ $\mu molkg^{-1}$.

Na mesma altitude do Océano Pacífico, especialmente nas rexións ecuatoriais, observamos que as masas de augas mostran concentracións baixas de $42 - 44$ $\mu molkg^{-1}$ na maior parte do Pacífico, influídas tamén pola chegada de correntes procedentes do Antártico como a AAIW. Esta tendencia seguirá a medida que aumente a profundidade, chegando a ter concentracións baixas e estables de DOC pola aparición da corrente profunda PDW cunhas cantidades moi baixas, como foi comentado a partir da Figura 5.7. Aínda así, seguen quedando zonas subtropicais cunhas concentracións algo máis elevadas, coma no caso do Pacífico ou coma no Índico, onde a rexión arredor dos $-30°$N de latitude presenta unhas concentracións de 46 $\mu molkg^{-1}$ que tamén aparecen representadas na gráfica de Roshan e DeVries (2017) ou nas figuras do artigo de Hansell et al. (2009). Para rematar, resaltar a acumulación bastante elevada de DOC atopada na costa de Sudáfrica ou no Mar Arábigo, sendo unhas estimacións algo máis elevadas que as presentadas por Roshan e DeVries (2017), cunhas concentracións próximas aos 50 $\mu molkg^{-1}$.

5.2.2 Seccións verticais do océano

Ademais das climatoloxías globais da predición do DOC descritas anteriormente, podemos representar graficamente seccións verticais do océano en diferentes lonxitudes. Nestas seccións verticais oceánicas móstranse as predicións do DOC con respecto á profundidade e tamén ao longo de distintas latitudes para un mes do ano en concreto. Estas representacións permiten interpretar a distribución das concentracións da variable DOC e a súa relación coas correntes oceánicas nas diferentes profundidades do océano.

Habitualmente, durante as viaxes oceánicas de investigación seleccionan rexións concretas onde realizar as mostras das distintas variables físicoquímicas do océano global. Polo tanto, estas seccións do océano con concentracións do DOC medidas *in situ* poden ser moi eficaces para valorar a precisión e adecuación das nosas predicións á realidade. Por este motivo, a continuación, imos presentar distintas seccións verticais oceánicas coñecidas para ser comparadas coa bibliografía existente ata o momento.

En primeiro lugar, na Figura 5.9 móstranse as predicións do DOC para a sección vertical do Océano Atlántico A16, lonxitude -29, 5°E, durante o mes de agosto. Os resultados aseméllanse bastante aos presentados polo artigo de Carlson et al. (2010), captando adecuadamente as diferentes concentracións das profundidades oceánicas. Na zona do Atlántico Norte (60°N-40°N) mostra concentracións máis elevadas, arredor dos $45 - 46$ $\mu molkg^{-1}$ a partir dos $500 - 1.000$ metros, que o resto de rexións do océano. Isto correspóndese coa rexión de afundimento das masas de auga superficiais procedentes das zonas subtropicais cargadas de DOC, xerando a corrente profunda do Atlántico Norte (NADW). Nas latitudes máis elevadas, as concentracións chegan, incluso, a uns valores de $48 - 50$ $\mu molkg^{-1}$. No resto do Océano Atlántico, en rexións profundas próximas ao ecuador, atopamos concentracións máis homoxéneas e constantes de, aproximadamente, 41 $\mu molkg^{-1}$.

A seguinte Figura 5.10 corresponde á sección vertical A20 situada nos 51, 5°E de lonxitude durante o mes de outubro. Esta rexión, situada entre os 45°N e os 5°N de latitude, abrangue dende a costa sueste de América do Norte ata a costa nordeste de América do Sur. As predicións obtidas tamén se parecen aos resultados empíricos reflectidos por Carlson et al. (2010), captando correctamente as maiores concentracións de DOC na rexións de maior latitude do océano debido á presenza da corrente profunda do Atlántico Norte (NADW) que resulta nunhas concentracións de $44 - 45$ $\mu molkg^{-1}$ arredor dos 1.500, 2.000 e 2.500 metros. No resto da rexión oceánica as concentracións mantéñense sobre os

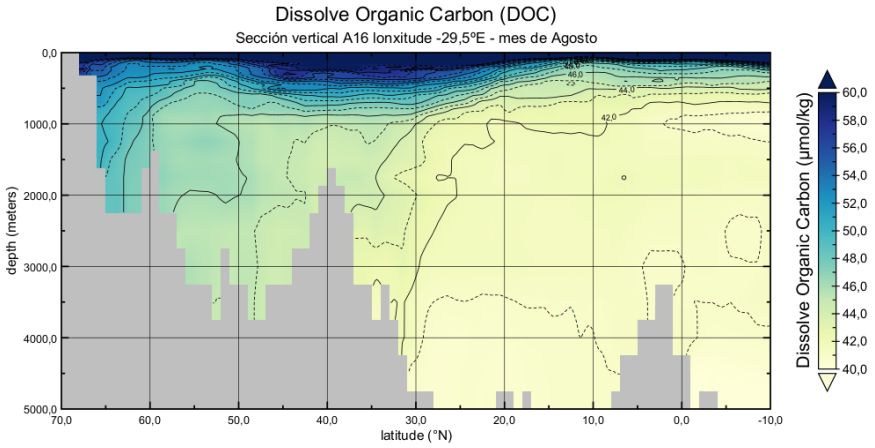

Figura 5.9: Sección vertical correspondente á sección A16 do Océano Atlántico na lonxitude $-29,5°$E durante o mes de Agosto.

42 $\mu molkg^{-1}$ en profundidades maiores aos 1.000 metros. Tal e como foi comentado nas anteriores climatoloxías, en profundidades próximas á superficie e incluso sobre os 500 metros de profundidade, as concentracións rondan os $55-60$ $\mu molkg^{-1}$ do mesmo xeito que aparece nos gráficos de Carlson et al. (2010) que utilizamos como referencia comparativa.

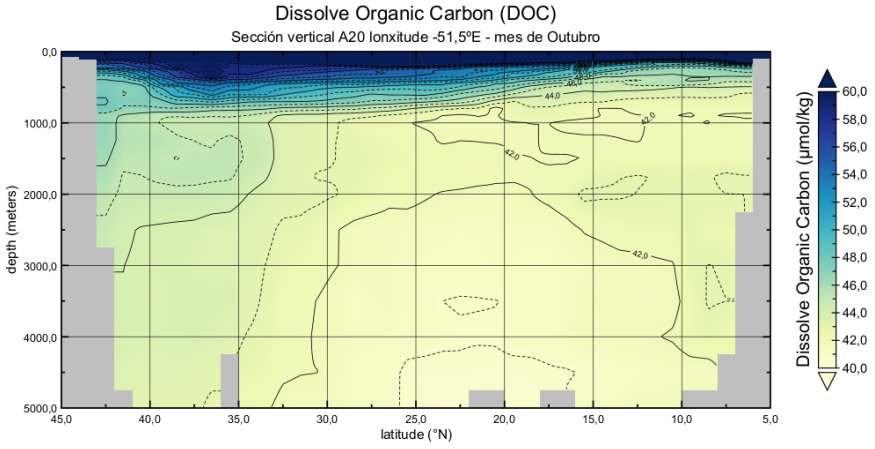

Figura 5.10: Sección vertical correspondente á sección A20 do Océano Atlántico na lonxitude $-51,5°$E durante o mes de Outubro.

Por último, foi seleccionada a sección A22 CLIVAR do Atlántico relativa á lonxitude de $-67,5°$E durante o mes de novembro (Figura 5.11). Esta selección mensual tamén corresponde con algúns dos meses onde se produciu esta toma de mostras empírica nesta rexión do océano. As predicións presentadas na Figura 5.11 mostran resultados bastante semellantes á sección A22 publicada por Carlson et al. (2010). A concentración do DOC arredor dos 1.000 e 2.000 metros, nas rexións por encima dos $-26°$N de latitude, atópase arredor dos $44-45$ $\mu molkg^{-1}$ e nas latitudes máis baixas, sobre os $-14°$N, sitúase sobre os 41 $\mu molkg^{-1}$ a partir desas profundidades; replicando con moita fiabilidade os resultados empíricos de Carlson et al. (2010) e, polo tanto, constatando que o modelo se asemella

con bastante precisión á realidade.

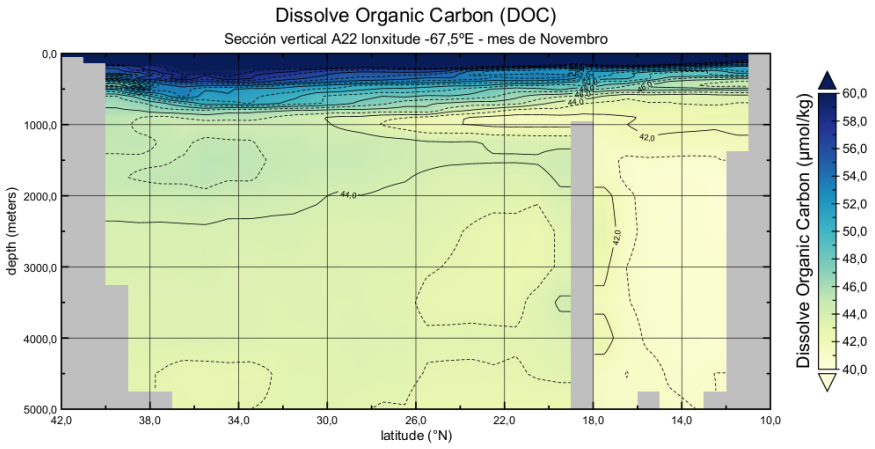

Figura 5.11: Sección vertical correspondente á sección A22 do Océano Atlántico na lonxitude $-67.5°$E durante o mes de Novembro.

Capítulo 6

Conclusións e traballo futuro

Para concluír, tras a presentación dos resultados obtidos co modelo axustado, expóñense as conclusións derivadas do estudo. Ademais, analízanse as posibles melloras e propóñense futuras liñas de investigación que poderían derivarse deste traballo no ámbito da oceanografía.

6.1 Conclusións

Neste traballo exploráronse diferentes metodoloxías para predicir as concentracións do carbono orgánico disolto (DOC) no océano global a partir dun conxunto de variables físicoquímicas. Inicialmente, axustáronse diversos modelos de aprendizaxe automática, entre os que se inclúen o Modelo Aditivo Xeneralizado (GAM), *Random Forest*, *Extreme Gradient Boosting* e o método dos k-veciños máis próximos (k-NN). A selección do modelo final realizouse en base a minimizar distintas funcións de perda (R^2, MAE, $RMSE$), determinando que o modelo máis axeitado para o obxectivo do estudo é o *Random Forest*.

A predición do DOC en función de outras variables físicoquímicas é unha aproximación coñecida neste campo, cos traballos previos de Roshan e DeVries (2017) ou Bonelli et al. (2022). No presente estudo, ademais, incorpórase unha compoñente mensual, seguindo a liña de investigación de Laine et al. (2024), quen aplicou esta estratexia para a estimación do DOC unicamente na capa superficial do océano. Por outra banda, e de maneira máis recente, Panaïotis et al. (2025) propuxeron unha aproximación baseada nun modelo estatístico alternativo, no que se incluían distintas profundidades para a estimación desta compoñente do ciclo do carbono. Con todo, o modelo desenvolvido no presente traballo amplía de maneira significativa esta proposta, ao considerar 40 niveis de profundidade fronte aos 4 empregados no citado estudo, o que permite unha caracterización máis detallada da distribución vertical do DOC. Deste xeito, as predicións que foron descritas durante a Sección 5.2 do Capítulo 5, permiten coñecer as concentracións da variable DOC ao longo de todo o océano global, nos distintos meses do ano e nas 40 profundidades elixidas. Porén, a metodoloxía elaborada neste estudo manexa unha cantidade de observacións e unha complexidade moito maior que calquera traballo realizado ata agora, converténdose nunha implementación innovadora neste campo.

De maneira máis específica, as predicións obtidas polo *Random Forest* proporcionan aproximacións das concentracións do DOC nas diferentes rexións do océano e ao longo das diferentes profundidades que resultan moi semellantes aos valores atopados de maneira empírica. O modelo predí adecuadamente as rexións superficiais, con maiores concentracións de DOC, como pode ser a rexión Ártica, as desembocaduras de grandes ríos ou zonas intermareais, e tamén detecta adecuadamente as baixas concentracións características do Océano Austral. Ademais, en maiores profundidades, o modelo recolle de maneira precisa o movemento das diferentes masas de auga e as correntes termohalinas que circulan

polos diferentes océanos do planeta. Así mesmo, describe con precisión o proceso de afundimento das masas de augas superficiais no norte do Océano Atlántico xunto co seu posterior descenso cara as zonas subtropicais, contribuíndo ao incremento das concentracións en profundidade nesta rexión. A pesar de certas diferenzas cuantitativas puntuais, as climatoloxías obtidas serven como unha aproximación moi salientable e verídica das concentracións deste composto do ciclo do carbono.

Pola contra, tamén é importante destacar que existen rexións do océano onde as predicións do modelo non recollen de maneira adecuada a distribución do DOC. Especialmente destacable é o que sucede na rexión norte do Océano Pacífico, a partir dos 1500 metros de profundidade. Nesta zona, territorio de múltiples viaxes oceánicas de investigación, o modelo presenta certas limitacións en canto a continuidade e suavidade das predicións. En particular, detectáronse cambios abruptos nas concentracións preditas que dan lugar a puntos con valores inxustificadamente elevados, cando os rexistros desta rexión do Pacífico non deberían ser tan altos. Ademais, esta situación atópase na rexión próxima aos $-180°$E de lonxitude, entre os $-30°$N e 30 °N de latitude, o que pode indicar a ausencia de suficientes datos no proceso de adestramento. Isto só é apreciable a partir de certa profundidade, polo que sería necesaria unha análise máis completa dos posibles motivos destes resultados.

Por último, a metodoloxía implementada neste traballo pode facilitar en gran medida a interpretación e presentación do carbono orgánico disolto (DOC). As predicións do océano global en diferentes profundidades e ao longo de distintos meses permiten: por un lado, presentar diferentes climatoloxías do océano global en superficie e nas diferentes profundidades; por outra banda, elaborar calquera sección vertical do océano para analizar a variación de concentracións coa profundidade e coa latitude; e, por último, comparar ambos gráficos mencionados ao longo dos distintos meses do ano coas correspondentes variacións estacionais das correntes oceánicas.

Como conclusión, o axuste e implementación do modelo de regresión para estimar e predicir o DOC presentado neste traballo, aínda que non é unha novidade dentro da bibliografía, aporta unha visión innovadora e moito máis completa ca os traballos previos. A pesar da necesidade de mellorar algunhas das predicións obtidas, esta metodoloxía pode dar un gran salto no estudo do DOC, do ciclo do carbono e, de forma máis ampla, do cambio climático e os seus efectos. Deste xeito, este traballo pretende avanzar un pequeno paso no camiño do coñecemento e ser, tamén, a semente de futuros traballos de investigación.

6.2 Liñas futuras

O seguinte paso na investigación das metodoloxías de estimación e aplicación de modelos de regresión, como continuación a este traballo, pode darse cara diferentes vías. En primeiro lugar, poderíanse axustar outros modelos de aprendizaxe automática que non foron considerados nesta investigación, pero si noutras da bibliografía, como as redes neuronais artificiais (ANN) ou os *Boosted Regression Trees* (BRTs).

Por outro lado, durante o proceso de adestramento e validación dos modelos, poderíase considerar a incorporación de bases de datos empíricas adicionais, máis completas e actualizadas, coa intención de ampliar o volume de información dispoñible para o axuste dos modelos de regresión. Esta ampliación permitiría, en principio, mellorar a precisión das predicións da variable resposta. Do mesmo xeito, resultaría conveniente avaliar a posible inclusión de outras variables físicoquímicas, recollidas no océano, co fin de enriquecer o conxunto de preditores e mellorar así o adestramento dos modelos. Neste sentido, a base de datos WOA 2023 ofrece variables complementarias de interese, como a porcentaxe de saturación de osíxeno (%) e a utilización aparente de osíxeno ($\mu molkg^{-1}$), que poderían ser integradas no proceso de modelaxe.

Por último, podería ser interesante afrontar o estudo e análise da aplicación deste ou doutros modelos estatísticos a contextos máis reducidos. Propoñer, seguindo a mesma metodoloxía, a estimación desta

variable do DOC, nunha rexión máis local do océano como podería ser a Ría de Vigo ou a costa galega; empregando unha base de datos limitada a esta rexión de interese.

Bibliografía

[1] Arıstegui, J., Duarte, C.M., Agustı, S., Doval, M., Álvarez-Salgado, X.A., & Hansell, D.A. (2002) Dissolved organic carbon support of respiration in the dark ocean. Science 298(5600): 1967-1967.

[2] Bercovici, S.K., Dittmar, T. & Niggemann, J. (2023) Processes in the Surface Ocean Regulate Dissolved Organic Matter Distributions in the Deep. Global Biogeochemical Cycles 37(12): e2023GB007740.

[3] Bonelli, A.G., Loisel, H., Jorge, D.S., Mangin, A., d'Andon, O.F., & Vantrepotte, V. (2022) A new method to estimate the dissolved organic carbon concentration from remote sensing in the global open ocean. Remote Sensing of Environment 281: 113227.

[4] Breiman, L., Friedman, J., Olshen, R.A., & Stone, C.J. (1984) Classification and regression trees. Chapman and Hall/CRC.

[5] Breiman, L., & Friedman, J.H. (1985) Estimating optimal transformations for multiple regression and correlation. Journal of the American statistical Association 80(391): 580-598.

[6] Breiman, L. (1996) Bagging predictors. Machine learning, 24: 123-140.

[7] Breiman, L. (2001) Random Forests. Machine Learning 45: 5-32.

[8] Broullón, D., Pérez, F.F., Velo Lanchas, A., Hoppema, M., Olsen, A., Takahashi, T., Key, R.M., Tanhua, T., Santana-Casiano, J.M. & Kozyr, A. (2020) A global monthly climatology of oceanic total dissolved inorganic carbon: A neural network approach. Earth System Science Data Discussions 2020: 1-30.

[9] Carlson, C.A., Hansell, D.A., Nelson, N.B., Siegel, D.A., Smethie, W.M., Khatiwala, S., Meyers, M.M. & Halewood E. (2010) Dissolved organic carbon export and subsequent remineralization in the mesopelagic and bathypelagic realms of the North Atlantic basin. Deep Sea Research Part II: Topical Studies in Oceanography 57(16): 1433-1445.

[10] Chen, T. & Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. En: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, New York, NY, USA, pp 785–794.

[11] De Boor, C. (1978) A practical guide to splines. Springer, New York.

[12] Dittmar, T., Lennartz, S.T., Buck-Wiese, H., Hansell, D.A., Santinelli, C., Vanni, C., Blasius, B. & Hehemann J.H. (2021) Enigmatic persistence of dissolved organic matter in the ocean. Nature Reviews Earth & Environment 2(8): 570-583.

[13] Drucker, H., Burges, C.J., Kaufman, L., Smola, A.J. & Vapnik, V. (1996) Support vector regression machines. Advances in Neural Information Processing Systems 9.

[14] Engel, A., Händel, N., Wohlers, J., Lunau, M., Grossart, H.P., Sommer, U., & Riebesell, U. (2011) Effects of sea surface warming on the production and composition of dissolved organic matter during phytoplankton blooms: results from a mesocosm study. Journal of Plankton Research 33(3): 357-372.

[15] Falkowski, P.G., & Raven, J.A. (2013) Aquatic photosynthesis. Princeton University Press.

[16] Fan, J., Gijbels, I., Hu, T.C., & Huang, L.S. (1996) A study of variable bandwidth selection for local polynomial regression. Statistica Sinica: 113-127.

[17] Fontela, M., García-Ibáñez, M.I., Hansell, D.A., Mercier, H., & Pérez, F.F. (2016) Dissolved organic carbon in the North Atlantic meridional overturning circulation. Scientific reports 6(1): 26931.

[18] Fontela, M., Pérez, F.F., Mercier, H. & Lherminier P. (2020) North Atlantic Western Boundary Currents Are Intense Dissolved Organic Carbon Streams. Frontiers in Marine Science 7:593757.

[19] Freund, Y., & Schapire, R. E. (1995) A desicion-theoretic generalization of on-line learning and an application to boosting. En: European conference on computational learning theory. Springer, pp 23-37.

[20] Friedman, H.J. (2001) Greedy Function Approximation: A Gradient Boosting Machine. Annals of Stadistics: 1189-1232.

[21] Greenwell, M.B. (2017). pdp: An R Package for Constructing Partial Dependence Plots. R package version 0.8.2. https://cran.r-project.org/web/packages/pdp/index.html

[22] Hansell, D.A., Carlson, C.A., Repeta, D.J., & Schlitzer, R. (2009) Dissolved organic matter in the ocean: A controversy stimulates new insights. Oceanography 22(4): 202-211.

[23] Hansell, D.A. (2013) Recalcitrant dissolved organic carbon fractions. Annual review of marine science 5(1): 421-445.

[24] Hansell, D.A. & Orellana, M.V. (2021) Dissolved Organic Matter in the Global Ocean: A Primer. Gels 7(3): 128.

[25] Hansell, D.A., Carlson, C.A., Amon, R.M.W., Álvarez-Salgado, X.A., Yamashita, Y., Romera-Castillo, C., Bif, M.B. (2021) Compilation of dissolved organic matter (DOM) data obtained from global ocean observations from 1994 to 2021, Version 2 (NCEI Accession 0227166). https://odv.awi.de/data/ocean/dom-compilation-hansell-et-al-2021-v2/. Acceso 15/11/2024.

[26] Hansell, D. A., Romera-Castillo, C., & Lopez, C.N. (2024) Dynamics of dissolved organic carbon in the global ocean. En: Hansell, D.A., & Carlson, C.A. (Third Edition) Biogeochemistry of Marine Dissolved Organic Matter. Elsevier, pp 769-802

[27] Hastie, T., & Tibshirani, R. (1990) Exploring the nature of covariate effects in the proportional hazards model. Biometrics: 1005-1016.

[28] Hastie, T., Tibshirani, R., & Friedman, J. H. (2017) The elements of statistical learning: Data mining, inference, and prediction. Springer, New York.

[29] Jiang, L., Pierrot, D., Wanninkhof, R., Feely, R.A., Tilbrook, B., Alin, S., Barbero, L., Byrne, R.H., Carter, B.R., Dickson, A.G., Gattuso, J., Greeley, D., Hoppema, M., Humphreys, M.P., Karstensen, J., Lange, N., Lauvset, S.K., Lewis, E.R., Olsen, A., Pérez, F.F., Sabine, C., Sharp, J.D., Tanhua, T., Trull, T.W., Velo, A., Allegra, A.J., Barker, P., Burger, E., Cai, W., Chen, C.A., Cross, J., Garcia, H., Hernandez-Ayon, J.M., Hu, X., Kozyr, A., Langdon, C., Lee, K., Salisbury, J., Wang, Z.A. & Xue, L. (2022) Best Practice Data Standards for Discrete Chemical Oceanographic Observations. Frontiers in Marine Science 8:705638.

[30] Jin, R., Gnanadesikan, A., & Holder, C. (2024). Using random forests to compare the sensitivity of observed particulate inorganic and particulate organic carbon to environmental conditions. Geophysical Research Letters 51(18): e2024GL110972.

[31] Kramer, O. (2013) K-Nearest Neighbors. En: Dimensionality Reduction with Unsupervised Nearest Neighbors. Intelligent Systems Reference Library, vol 51. Springer, Berlin, Heidelberg.

[32] Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C. & Hunt, T. (2024) caret: Classification and Regression Training. R package version 7.0-1. https://cran.rproject.org/web/packages/caret/index.html.

[33] Laine, M., Kulk, G., Jönsson, B.F., & Sathyendranath, S. (2024) A machine learning model-based satellite data record of dissolved organic carbon concentration in surface waters of the global open ocean. Frontiers in Marine Science 11: 1305050.

[34] Lang, S., & Brezger, A. (2004) Bayesian P-splines. Journal of computational and graphical statistics 13(1): 183-212.

[35] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (2002) Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11): 2278-2324.

[36] LeCun, Y., Bengio, Y., & Hinton, G. (2015) Deep learning. nature 521(7553): 436-444.

[37] Lee, T.R., Wood, W.T., & Phrampus, B.J. (2019) A machine learning (kNN) approach to predicting global seafloor total organic carbon. Global Biogeochemical Cycles 33(1): 37-46.

[38] Li, C., Wu, H., Yang, C., Cui, L., Ma, Z., & Wang, L. (2024) Advanced Machine Learning Models for Estimating the Distribution of Sea-Surface Particulate Organic Carbon (POC) Concentrations Using Satellite Remote Sensing Data: The Mediterranean as an Example. Sensors 24(17): 5669.

[39] Liu, H. (2008) Generalized additive model. Department of Mathematics and Statistics University of Minnesota Duluth: Duluth, MN, USA 55812.

[40] Liu, H., Li, Q., Bai, Y., Yang, C., Wang, J., Zhou, Q., Hu, S., Shi, T., Liao, X & Wu, G. (2021) Improving satellite retrieval of oceanic particulate organic carbon concentrations using machine learning methods. Remote Sensing of Environment 265: 112316.

[41] Lønborg, C., Carreira, C., Jickells, T., & Álvarez-Salgado, X.A. (2020) Impacts of global change on ocean dissolved organic carbon (DOC) cycling. Frontiers in Marine Science 7: 466.

[42] Lu, B., & Hardin, J. (2021) A unified framework for random forest prediction error estimation. Journal of Machine Learning Research 22(8): 1-41.

[43] McCullough, P. & Nelder, J.A. (1989) Generalized Linear Models, 2nd ed. Chapman and Hall/CRC.

[44] McCulloch, W.S., & Pitts, W. (1990) A logical calculus of the ideas immanent in nervous activity. Bulletin of mathematical biology 52: 99-115.

[45] Myers, R.H. & Montgomery, D.C. (1997) A Tutorial on Generalized Linear Models. Journal of Quality Technology, 29(3): 274–291.

[46] NASA Goddard Institute for Space Studies (2023) Panoply Data Viewer. Version 5.5.5. https://www.giss.nasa.gov/tools/panoply. Acceso 21 de Xaneiro 2025.

[47] Natekin, A. & Knoll, A. (2013) Gradient boosting machines, a tutorial. Frontiers in neurorobotics 7:21.

[48] Nelder, J.A. & Wedderburn, R.W.M. (1972) Generalized Linear Models. Journal of the Royal Statistical Society Series A: Statistics in Society 135(3): 370–384

[49] Panaïotis, T., Wilson, J., & Cael, B. (2025) A machine learning-based dissolved organic carbon climatology. Geophysical Research Letters 52(7): e2024GL112792.

[50] R Core Team (2025) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. R project version 4.5.0 https://www.R-project.org/.

[51] Reagan, J.R., Boyer, T.P., García, H.E., Locarnini, R.A., Baranova, O.K., Bouchard, C., Cross, S.L., Mishonov, A.V., Paver, C.R., Seidov, D., Wang, Z., Dukhovskoy, D. (2024) World Ocean Atlas 2023. NOAA National Centers for Environmental Information. https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.nodc:0270533. Acceso 29/01/2025.

[52] Romera-Castillo, C., Álvarez, M., Pelegrí, J.L., Hansell, D.A., & Álvarez-Salgado, X.A. (2019) Net additions of recalcitrant dissolved organic carbon in the deep Atlantic Ocean. Global Biogeochemical Cycles 33(9): 1162-1173.

[53] Roshan, S., & DeVries, T. (2017) Efficient dissolved organic carbon production and export in the oligotrophic ocean. Nature communications 8(1): 2036.

[54] Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1986) Learning representations by back-propagating errors. nature 323(6088): 533-536.

[55] Schapire, R.E. (2003) The boosting approach to machine learning: An overview. En: Denison, D.D., Hansen, M.H., Holmes, C.C., Mallick, B., & Yu, B. (Eds.) Nonlinear estimation and classification, vol 171. Springer, New York, NY: 149-171.

[56] Schlitzer, R. (2002) Carbon export fluxes in the Southern Ocean: Results from inverse modeling and comparison with satellite based estimates. Deep-Sea Research Part II: Topical Studies in Oceanography 49(9-10): 1623–1644.

[57] Schlitzer, R. (2007) Assimilation of radiocarbon and chlorofluorocarbon data to constrain deep and bottom water transports in the world ocean. Journal of Physical Oceanography 37(2): 259–276.

[58] Song, Y., Liang, J., Lu, J., & Zhao, X. (2017) An efficient instance selection algorithm for k nearest neighbor regression. Neurocomputing 251: 26-34.

[59] Specht, D.F. (1991) A general regression neural network. IEEE transactions on neural networks 2(6): 568-576.

[60] Thornton, D.C. (2014) Dissolved organic matter (DOM) release by phytoplankton in the contemporary and future ocean. European Journal of Phycology 49(1): 20-46.

[61] Toggweiler, J.R., & Key, R.M. (2001) Thermohaline circulation. Encyclopedia of ocean sciences 6: 2941-2945.

[62] Verdugo, P., Alldredge, A.L., Azam, F., Kirchman, D.L., Passow, U., & Santschi, P.H. (2004) The oceanic gel phase: a bridge in the DOM–POM continuum. Marine chemistry 92(1-4): 67-85.

[63] Wand, M.P., & Jones, M.C. (1994) Kernel smoothing. Chapman & Hall/CRC.

[64] Watson, R.T. (2001) Cambio climático 2001: Informe de síntesis. Contribución de los Grupos de Trabajo I, II, y III al Tercer Informe de Evaluación del Grupo Intergubernamental de Expertos sobre el Cambio Climático. IPCC, 1-206. https://www.ipcc.ch/site/assets/uploads/2018/08/TARsyrfulles.pdf.

[65] Wood, S.N. (2017) Generalized Additive Models: An Introduction with R (2nd edition). Chapman and Hall/CRC.

[66] Wright, M.N., Wager, S. & Probst, P. (2024) ranger: A Fast Implementation of Random Forests. R package version 0.17.0. https://cran.r-project.org/web/packages/ranger/index.html. Acceso 1 Decembro de 2024.

[67] Wu, H., Cui, L., Wang, L., Sun, R., & Zheng, Z. (2023) A method for estimating particulate organic carbon at the sea surface based on geodetector and machine learning. Frontiers in Marine Science 10: 1295874.

[68] Xiang, D. (2001) Fitting generalized additive models with the GAM procedure. En: SUGI Proceedings. Cary NC: SAS Institute, Inc., pp 256-326.

[69] Yang, C., Kim, T., Wang, R., Peng, H., & Kuo, C.C.J. (2019) Show, attend, and translate: Unsupervised image translation with self-regularization and attention. IEEE Transactions on Image Processing 28(10): 4845-4856.

[70] Ying, C., Qi-Guang, M., Jia-Chen, L., & Lin, G. (2013) Advance and prospects of AdaBoost algorithm. Acta Automatica Sinica 39(6): 745-758.

[71] Zhang, Z. (2016) Introduction to machine learning: k-nearest neighbors. Annals of translational medicine 4(11): 218.

Índice de táboas

Índice de figuras