

Assessing the Impact of Dataset Shift for Automatic *Nephrops norvegicus* Burrows Detection and Classification Through Deep Learning

Oscar Papini¹, Marco Reggiannini¹, Enrico Cecapolli², Filippo Domenichetti², Lorenzo Zacchetti^{2,3}, Michela Martinelli², and Gabriele Pieri¹

¹ Institute of Information Science and Technologies (ISTI), National Research Council of Italy, Via G. Moruzzi 1, 56124 Pisa, Italy

{oscar.papini,marco.reggiannini,gabriele.pieri}@isti.cnr.it

² Institute for Marine Biological Resources and Biotechnology (IRBIM), National Research Council of Italy, Largo Fiera della Pesca 2, 60125 Ancona, Italy

{enrico.cecapolli,lorenzo.zacchetti}@irbim.cnr.it

{filippo.domenichetti,michela.martinelli}@cnr.it

³ Department of Biological, Geological, and Environmental Sciences (BiGeA), Alma Mater Studiorum – University of Bologna, Piazza di Porta S. Donato 1, 40126 Bologna, Italy

Abstract In this work, machine learning and computer vision methods are investigated to support the estimation of the *Nephrops norvegicus* (Norway lobster) population, a key task for sustainable fisheries management and marine ecosystem monitoring. This task is currently carried out by manually counting the animals' burrows that are visually recognisable in video surveys captured in the area of interest, and then using those counts to estimate the species density in the area. To reduce uncertainty in human assessment, a burrow classifier based on a supervised learning model has been developed to be integrated into the pipeline. A central challenge in deploying such models in real-world marine environments is dataset shift, arising from variations in both acquisition conditions and ecological contexts. Indeed, several factors, including water turbidity, irregular lighting, and alterations of colour, potentially hinder the accuracy of the detection. Furthermore, geographic and temporal variability, such as differences between survey campaigns or regions, can exacerbate these shifts, leading to reduced generalisation. This work concerns the application of the proposed classifier to data acquired in two different areas: the Adriatic Sea and the northern Atlantic Ocean, off the coast of Ireland. As shown in the paper, the differences in conformation and species type between those two areas greatly affect the performances of the classifier. Nevertheless, the proposed method can be employed in the entire range of the species distribution, provided that the model is properly fine-tuned for the specific geomorphological features of the area in which it is applied.

Keywords: Dataset shift · *Nephrops norvegicus* · Underwater Television surveys · Deep Learning

1 Introduction

Nephrops norvegicus (Norway lobster) is a burrowing crustacean inhabiting the continental shelf of the northeastern Atlantic Ocean and the Mediterranean Sea, at typical depths ranging between 20 and 800 metres [1]. It is among the most commercially relevant species in Europe, with landings amounting to about 60 000 tonnes per year, even though the trend seems to be decreasing in recent years [2].

Towed Underwater Television (UWTV) surveys are one of the most accredited methods conventionally adopted by many European countries for stock assessment purposes, as they allow estimating the burrows density (number of burrows per square metre), which is considered a proxy for the population abundance [3]. The estimation is provided by trained analysts who visually inspect hours of underwater video footage of a known seabed area to recognise and count the number of *N. norvegicus* burrows. The inspection of a video is usually performed independently by multiple operators, and the outcome is returned as the average of the individual measurements, which is eventually employed to estimate the burrows density.

Since this procedure relies on the direct assessment of human operators, the outcome is prone to subjectivity and uncertainty errors. To attenuate these factors, the International Council for the Exploration of the Sea (ICES) issues specific guidelines and promotes training workshops intended to standardise the video reading methodologies [4]. Despite these efforts of the international scientific community, the counting task keeps being an entirely manual and time-consuming process, as well as highly sensitive to human errors [5]. Undoubtedly, an automatic tool that is able to detect and count *N. norvegicus* burrows would be invaluable, both in terms of manual labour relief and of reduced subjectivity. Nonetheless, this is a challenging task even for state-of-the-art object detection models such as YOLO [6].

Among the issues that an automatic detector of *N. norvegicus* burrows can run into, the vast environmental diversity that can be found in the *N. norvegicus* habitat is one of the main ones. Indeed, the differences in the biological features of the animals, as well as in the geophysical properties of the sea floor, induce a strong differentiation in the visual characteristics of the burrow entrances, to the point that even expert human annotators may have difficulty recognising them in footage collected from geographic regions unfamiliar to them. Moreover, despite having to follow a common standard protocol, the equipment used by the various research institutions for the UWTV recording differs in terms of cameras setup and specifications [7].

In this work, the domain generalization capability of a deep learning-based *N. norvegicus* burrows classifier are investigated through the use of two datasets originating from two different geographic areas: the Pomo/Jabuka Pits area in

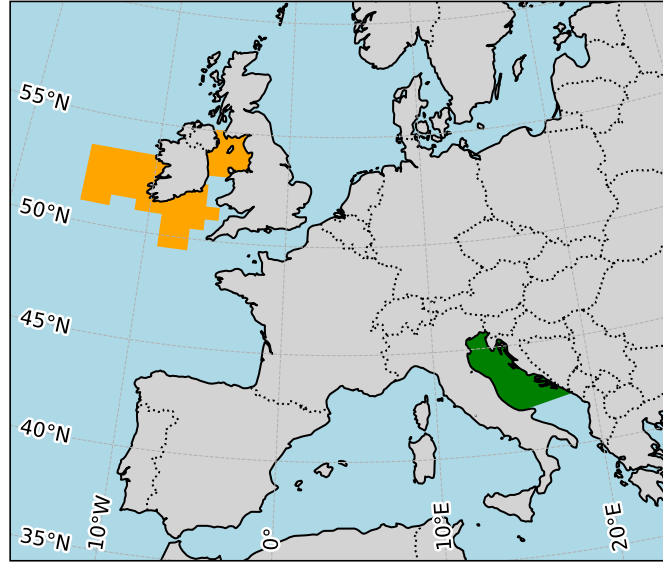


Figure 1: Acquisition zones for the two datasets (green: Adriatic dataset; orange: Irish dataset).

the Adriatic Sea (Central Mediterranean basin), which represents a Vulnerable Marine Ecosystem and an Essential Fish Habitat for various commercially and ecologically important species [8, 9], and the North Atlantic Ocean off the coast of Ireland (see Figure 1). The chosen deep learning model has been described in previous works by the same authors [6] and consists of the cascaded sequence of an encoding block, in charge of extracting features and providing a latent representation of the data, followed by a fully connected block that performs the actual recognition of the candidate burrow entrance. Those two blocks are trained independently to provide both within-region and cross-region evaluations for the classification task.

The remainder of this paper is organised as follows. Section 2 describes the two datasets that are used in this work; Section 3 presents the deep learning-based detection pipeline; Section 4 details the setup of the experiments; in Section 5 the results are reported and briefly discussed; finally, Section 6 concludes the paper and outlines future research directions.

2 Datasets

As mentioned in the Introduction, the distribution of *N. norvegicus* habitats involve marine environments which are greatly diverse in terms of geomorphological characteristics. Moreover, footage is captured by different institutions using different setups in terms of e.g. camera angle and illumination apparatus. These

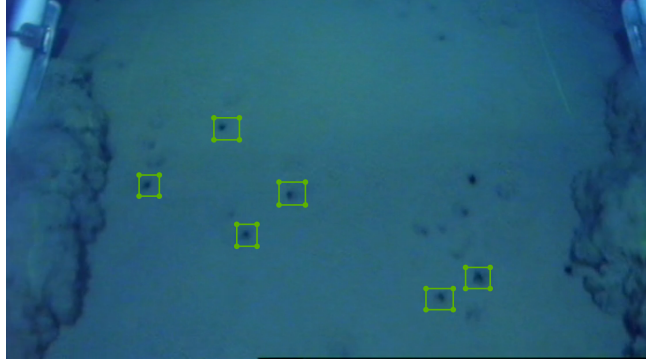


Figure 2: Example of an annotated frame in the Adriatic dataset. Green boxes denote NEPHROPS objects.

differences do not affect the traditional *N. norvegicus* burrows counting method, but automated computer vision models may learn the peculiarities of a specific institute or location instead of the relevant features of *N. norvegicus* burrows entrances.

In this study, two datasets acquired in different geographic domains are used to train the same deep learning architecture and assess the robustness of the resulting models.

2.1 Adriatic dataset

Starting from footage collected during surveys conducted by IRBIM-CNR in the Central Adriatic Sea, more precisely in the Pomo/Jabuka Pits area, a set of 1110 frames has been extracted and, following the guidelines in [10], the images were annotated by tracing a bounding box, labelled NEPHROPS (N), around objects recognised as entrances of *N. norvegicus* burrows. An example of an annotated frame is provided in Figure 2.

Subsequently, a custom automatic segmentation method has been used to identify other objects in the field of view that may be confused with *N. norvegicus* entrances, such as other species' entrances or roundish shadows; these objects have been automatically marked with bounding boxes with a dedicated NON-NEPHROPS (NN) label. In the end, only 793 images contain at least one bounding box, for a total of 717 NEPHROPS and 1240 NON-NEPHROPS objects, but all the 1110 frames are part of the final dataset.

2.2 Irish dataset

A dataset from footage collected during surveys in the Irish Sea, the Celtic Sea and the Northern Atlantic Ocean around Ireland has been published by the Irish Marine Institute [11]. It consists of 468 annotated images with labelled bounding boxes belonging to the following classes:

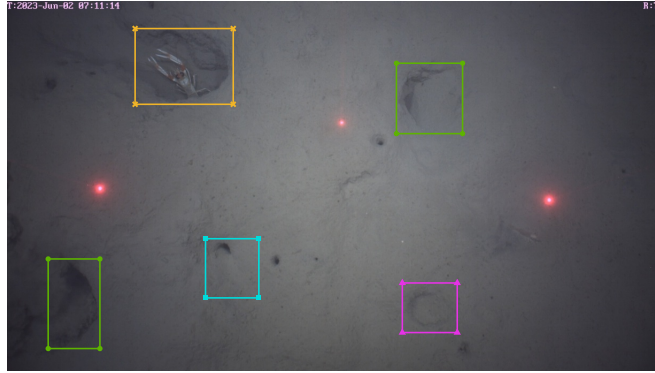


Figure 3: Example of an annotated frame in the Irish dataset. Green boxes with circle corners denote PRAWN_BURROW objects; cyan boxes with square corners denote SMALL_PRAWN_BURROW objects; magenta boxes with triangle corners denote CLOSED_BURROW objects; yellow boxes with cross corners denote GATEKEPT_BURROW objects.

- PRAWN_BURROW (P), corresponding to entrances associated with *N. norvegicus* burrows (872 bounding boxes);
- SMALL_PRAWN_BURROW (sP), also corresponding to entrances associated with *N. norvegicus* burrows, albeit smaller in size (548);
- GATEKEPT_BURROW (Gk), corresponding to *N. norvegicus* burrow entrances where a specimen of *N. norvegicus* (or a part of it, e.g. a pincer) is visible in the entrance (79);
- CLOSED_BURROW (Cl), corresponding to *N. norvegicus* entrances of abandoned burrows that have collapsed (173);
- CRAB_BURROW (Cr), corresponding to entrances associated with burrows of other species (99).

An example of an annotated frame is provided in Figure 3.

2.3 Tiled datasets

Most object detectors work at *image* level, that is to say, a whole frame is fed to the model and a list of candidate objects, in the form of bounding boxes, is returned. In this work, the problem of detecting *N. norvegicus* burrows entrances is tackled as a classification problem: given a portion of suitable dimensions (a *tile*) obtained as a crop of an image similar to the frames extracted from UWTV footage, determine whether it contains an entrance of a *N. norvegicus* burrow by giving a (global) label to the tile. To do so, two datasets of labelled square tiles have been derived from the Adriatic and the Irish datasets through the following procedure:

1. to make the two datasets comparable, images have been preprocessed to take into account the perspective distortion due to the camera inclination and different image acquisition sizes;
2. the images have been divided into 60×60 px tiles with 30 px overlap (this size was selected empirically to be comparable with the average area occupied by the entrances in the images);
3. each tile is given a label depending on the labels of the bounding boxes intersecting it, and an additional BACKGROUND (BG) label is given to tiles that do not intersect any bounding box.

In particular, three labels are possible for tiles from the Adriatic dataset (NEPHROPS, NON-NEPHROPS and BACKGROUND), and six labels for tiles from the Irish dataset (the five labels of [11], plus BACKGROUND).

3 Models

The pipeline outlined in this work consist of two separate phases: a feature extraction step followed by a classification step. The first step was performed using a type of neural networks known as *autoencoders* (AEs), whose main task is to learn an efficient representation, or *encoding*, of data, usually belonging to a lower-dimensional domain (sometimes called *latent space*). AEs are composed of an *encoder* and a *decoder*, the latter being the portion devoted to reconstruct the original data from the latent representation. Since AEs do not use labelled data, it is possible to train them without the involvement of the expert annotators.

The philosophy of the second step, instead, is to use the learned representation as the basis for the classification task—in other words, instead of using samples of the form (image, label) for the supervised training of the classification model, the samples have the form (encoding, label). Thus, the weights of the trained encoder are frozen, and any classifier can be chained to the encoder and trained. In this work, a simple multilayer perceptron (MLP), i.e., a stack of fully connected linear layers, is used as the classifying model.

In summary, the overall training process can be described as (see Figure 4):

1. train an AE using a large number of (unlabelled) tiles and freeze its weights;
2. train a classifier (e.g. a MLP) using as inputs the encodings of the labelled tiles with respect to the AE.

All the experiments have been carried out within a PyTorch (version 2.6.0) framework. Tables 1 and 2 detail the architectures of the networks involved in this study; in those tables, the output of a layer is fed directly to the next one. We remark that this paper represents mainly a proof of concept, and these architectures will be further analysed and improved in future works.

4 Training setup

To assess the effects of the different visual characteristics on the latent features, two different autoencoder models with the same architecture were trained independently, once using tiles from the Adriatic dataset and once using Irish

Table 1: List of autoencoder layers.

Layer	Parameters	Output dim.
Encoder		
(input)	–	$3 \times 60 \times 60$
CONV2D($C = 8, k = 3, s = 1, p = 0$)	224	$8 \times 58 \times 58$
LEAKYReLU	–	–
CONV2D($C = 16, k = 3, s = 2, p = 0$)	1168	$16 \times 28 \times 28$
LEAKYReLU	–	–
CONV2D($C = 32, k = 3, s = 2, p = 0$)	4640	$32 \times 13 \times 13$
LEAKYReLU	–	–
FLATTEN	–	5408
LINEAR($n = 5408, m = 1024$)	5 538 816	1024
LEAKYReLU	–	–
LINEAR($n = 1024, m = 128$)	131 200	128
LEAKYReLU	–	–
Decoder		
(input)	–	128
LINEAR($n = 128, m = 1024$)	132 096	1024
LEAKYReLU	–	–
LINEAR($n = 1024, m = 5408$)	5 543 200	5408
LEAKYReLU	–	–
UNFLATTEN	–	$32 \times 13 \times 13$
UPSAMPLE($n = 2$)	–	$32 \times 26 \times 26$
CONVTRANSPOSE2D($C = 16, k = 3, s = 1, p = 0$)	4624	$16 \times 28 \times 28$
LEAKYReLU	–	–
UPSAMPLE($n = 2$)	–	$16 \times 56 \times 56$
CONVTRANSPOSE2D($C = 8, k = 3, s = 1, p = 0$)	1160	$8 \times 58 \times 58$
LEAKYReLU	–	–
CONVTRANSPOSE2D($C = 3, k = 3, s = 1, p = 0$)	219	$3 \times 60 \times 60$
SIGMOID	–	–

Table 2: List of perceptron layers. n_c refers to the number of output classes ($n_c = 3$ for Adriatic dataset, $n_c = 6$ for Irish dataset).

Layer	Parameters	Output dim.
(input)	–	128
LINEAR($n = 128, m = 64$)	8256	64
LEAKYReLU	–	–
LINEAR($n = 64, m = 16$)	1040	16
LEAKYReLU	–	–
LINEAR($n = 16, m = n_c$)	$17 \cdot n_c$	n_c
SOFTMAX	–	–

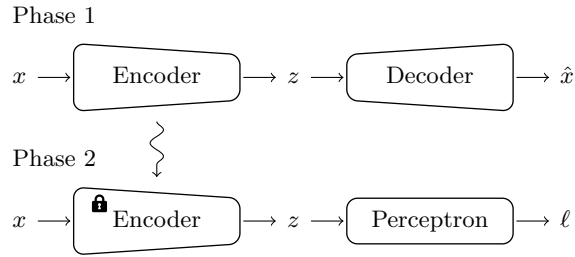


Figure 4: Concept of training phases. Here x is the original tile, z is the tile encoding, $\hat{x}$ is the reconstructed tile, and ℓ is the predicted label.

tiles. More in detail, in both cases a set of 16384 tiles was randomly selected from the relative dataset and used for training. To reduce the impact of almost uniform tiles, which bring little information, the standard deviation σ of the pixels intensity of each tile was computed, and the selection has been carried out so that a substantial part of the chosen tiles shows sufficient variability. The AEs were trained to maximise their tile reconstruction capability using the MSE loss, with a small L^1 loss component added to disfavour an average blurry reconstruction and preserve contrast. A further set of 4096 tiles, with the same variability distribution as the training set, was selected for each dataset to be used as a validation set for the implementation of an early stopping policy.

Perceptrons in the second phase were trained using the standard multi-class cross-entropy as the loss function. A stratified 5-fold cross-validation scheme is implemented: five sets of 768 tiles were extracted independently from both the Adriatic and the Irish datasets, so that a suitable number of samples are represented for each class. Using these sets, five models are trained with one of the folds acting as validation set for early stopping and the union of the others as training set. Another set of 480 tiles, with the same labels proportion as the training folds, is extracted from each dataset to be used as a test set after the training of each model. The tiles were selected at random while ensuring that all tiles extracted from the same frame were assigned to either the training set or the test set, to prevent potential overfitting due to spatial correlations.

To evaluate the transferability of the learned representation, all the four possible combinations of AE/MLP training were considered, namely (a) AE and MLP both trained on Adriatic tiles; (b) AE trained on Irish tiles, MLP trained on Adriatic tiles; (c) AE trained on Adriatic tiles, MLP trained on Irish tiles; (d) AE and MLP both trained on Irish tiles.

5 Results and discussion

Figure 5 shows some examples of reconstructions of tiles coming from different locations, allowing a qualitative evaluation of the autoencoders reconstruction effectiveness. The results of the perceptrons training experiments are shown

in Tables 3 and 4. In those tables, “AE/MLP Adriatic” (respectively, “Irish”) means “AE/MLP trained on the Adriatic dataset” (respectively, on the Irish dataset). The confusion matrices in Table 3 display the 5-fold average of the *test* sets classifications, where each cell (i, j) represents the number of tiles having ground truth (G.T.) label i predicted as label j . In Table 4, the capability of the MLPs to correctly identify *N. norvegicus* burrows entrances is assessed using as evaluation metrics the precision and recall computed for training, validation and test sets limited to the labels related with active *N. norvegicus* burrows (i.e., NEPHROPS for the Adriatic dataset; PRAWN_BURROW, SMALL_PRAWN_BURROW and GATEKEPT_BURROW for the Irish one, combining the three classes in the latter case).

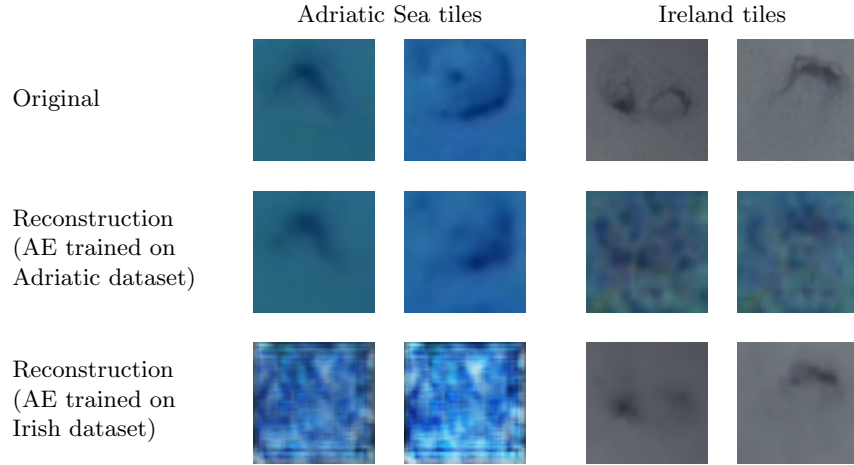


Figure 5: Examples of autoencoder reconstructions.

When evaluating the quality of reconstructions, AEs consistently demonstrate high performance when applied to tiles coming from the same geographic area as the dataset on which they have been trained, thus proving that they effectively learn the underlying distribution and features of that specific domain. However, quality clearly degrades in cross-dataset scenarios. Interestingly, the capability of the Adriatic-trained AE to reconstruct Irish data appears to be slightly superior to the one of the opposite combination. This discrepancy is likely attributable to the higher quality and potentially better-defined visual characteristics of the Irish images, which may offer a more robust structural signal, allowing the Adriatic model to extract a more informative representation of data despite the domain shift.

This behaviour is also reflected in the results of the tiles classification through MLPs. Indeed, classifiers trained on the Irish dataset achieve higher precision and recall than the ones trained on the Adriatic dataset, with an increase of

Table 3: Confusion matrices of MLP experiments computed on the 480 test tiles.

(a) AE Adriatic Sea, MLP Adriatic Sea				(b) AE Ireland, MLP Adriatic Sea			
G.T.	Prediction			G.T.	Prediction		
	BG	N	NN		BG	N	NN
BG	142.4	7	10.6	BG	111.2	21.2	27.6
N	5.4	106.2	48.4	N	38	85.4	36.6
NN	17.2	56	86.8	NN	60.4	60.4	39.2

(c) AE Adriatic Sea, MLP Ireland							(d) AE Ireland, MLP Ireland						
G.T.	Prediction						G.T.	Prediction					
	BG	P	sP	Gk	CL	CR		BG	P	sP	Gk	CL	CR
BG	153.4	15.8	18.6	0.4	1	0.8	BG	144.4	17.2	23.8	2.2	1.4	1
P	18	79.2	17	2	1.2	1.6	P	13	86.2	14.8	2.8	2.2	0
sP	36.4	27	51.8	0	1.2	0.6	sP	31.4	24.8	57.4	0.2	2.4	0.8
Gk	0.6	4	0	3	0	0.4	Gk	0.2	3.4	0	4.2	0	0.2
CL	14.8	8.6	6.2	0	1.4	0	CL	11.4	8.2	5.8	0	5.6	0
CR	7	2.8	3.6	0	0	1.6	CR	5.8	2.2	3.8	0.2	0	3

Table 4: Perceptron performances for *N. norvegicus*-related classes.

			Training		Validation		Test	
			<i>p</i>	<i>r</i>	<i>p</i>	<i>r</i>	<i>p</i>	<i>r</i>
(a)	Adriatic	Adriatic	0.703	0.679	0.625	0.605	0.630	0.664
(b)	Irish	Adriatic	0.549	0.571	0.531	0.551	0.513	0.534
(c)	Adriatic	Irish	0.798	0.757	0.759	0.718	0.769	0.754
(d)	Irish	Irish	0.821	0.841	0.755	0.765	0.754	0.794

All values are rounded to the third decimal digit.

12.4% in precision and 13% in recall on the respective test sets when considering models whose AE and MLP are trained on the same dataset (rows (a) and (d) of Table 4). The more robust feature representation in the encoding of Irish tiles influences the classification result also in cross-dataset cases (rows (b) and (c)). Nonetheless, within-dataset models show a more consistent classification effectiveness, thus highlighting the importance of area-specific models for each possible geographic location.

6 Conclusion

This work highlights the integration of machine learning techniques into *N. norvegicus* population monitoring as a vital step toward sustainable fisheries man-

agement. By automating burrow detection, we could reduce the subjectivity and manual labour inherent in current survey assessments, providing a more reliable foundation for density estimation.

However, our findings highlight that one of the main obstacles to large-scale deployment is dataset shift. As evidenced by differences in the model's performance across the two study areas, model accuracy is strongly influenced by site-specific factors, including turbidity, lighting and geomorphological diversity. These environmental discrepancies act as a barrier to generalisation and direct the research activities towards models that should not be treated as static tools but as an adaptive framework that requires fine-tuning tailored to the specific conditions of each surveyed region.

In the future, further analyses will be conducted, accounting for additional variations in the architectures used for both the AE and the MLP.

Acknowledgments. The authors would like to thank the IRBIM-CNR staff who contributed to collecting UWTV footage, in particular Andrea Belardinelli, Pierluigi Penna and Paolo Scarpini. The authors would also like to thank Giulio Del Corso (ISTI-CNR) for the inspirational conversations.

Funding. This work is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101000825 – NAUTILUS Project (<https://www.nautilus-h2020.eu>).

Data Availability. Due to strict regulations concerning both the preservation of fragile marine ecosystems and economic agreements concerning fishing, related sensitive information potentially contained in the Adriatic Sea dataset cannot be shared without the necessary preprocessing steps, e.g. data anonymisation. In the future, data and code will be released, provided that the mentioned procedures are applied.

Authors' contribution. Conceptualization: O.P., M.R., E.C., F.D., L.Z., M.M., G.P.; Data curation: E.C., F.D., L.Z., M.M.; Investigation: O.P., M.R., G.P.; Methodology: O.P., M.R., G.P.; Resources: O.P., M.R., E.C., F.D., L.Z., M.M., G.P.; Software: O.P., M.R., G.P.; Supervision: M.M., G.P.; Visualization: O.P.; Writing – original draft: O.P., M.R., E.C., F.D., L.Z., M.M., G.P.; Writing – review & editing: O.P., M.R., E.C., M.M., G.P. All authors have read and approved the final manuscript.

Disclosure of Interests. The authors of this work declare that they have no conflicts of interest.

References

1. Bell, M.C., Tuck, I., Dobby, H.: '*Nephrops* Species'. In: Lobsters: Biology, Management, Aquaculture and Fisheries. Ed. by B.F. Phillips. 2nd Edition. John Wiley & Sons, Ltd, 2013. Chap. 12, pp. 357–413. <https://doi.org/10.1002/9781118517444.ch12>.
2. Aguzzi, J. *et al.*: Advancing fishery-independent stock assessments for the Norway lobster (*Nephrops norvegicus*) with new monitoring technologies. *Front. Mar. Sci.* **9** (2022). <https://doi.org/10.3389/fmars.2022.969071>

3. Aguzzi, J. *et al.*: Burrow emergence rhythms of *Nephrops norvegicus* by UWTV and surveying biases. *Sci. Rep.* **11** (2021). <https://doi.org/10.1038/s41598-021-85240-3>
4. ICES, Working Group on *Nephrops* Surveys (WGNEPS; outputs from 2022 meeting). *ICES Sci. Rep.* 5:26, 125 pp. (2023). <https://doi.org/10.17895/ices.pub.22211161.v1>
5. ICES, Report of the Workshop on *Nephrops* Burrow Counting (WKNEPS). *ICES CM 2018/EOSG:25*, 44 pp. 2–5 October 2018, Aberdeen, UK (2019). <https://doi.org/10.17895/ices.pub.8180>
6. Papini, O. *et al.*: Deep Learning–Based Detection of *Nephrops norvegicus* Burrows. In: *Proceedings of the 2025 IEEE International Workshop on Metrology for the Sea*, pp. 39–44 (2025). <https://doi.org/10.1109/MetroSea66681.2025.11245690>
7. Dobby, H. *et al.*: ICES Survey Protocols – Manual for *Nephrops* Underwater TV Surveys, coordinated under ICES Working Group on *Nephrops* Surveys (WGNEPS). *ICES Tech. Mar. Environ. Sci.* **65** (2021). <https://doi.org/10.17895/ices.pub.8014>
8. Martinelli, M. *et al.*: Towed underwater television towards the quantification of Norway lobster, squat lobsters and sea pens in the Adriatic Sea. *Acta Adriat.* **54**(1), 3–12 (2013)
9. Martinelli, M. *et al.*: Changes in Abundance and Distribution of the Sea Pen, *Funiculina quadrangularis*, in the Central Adriatic Sea (Mediterranean Basin) in Response to Variations in Trawling Intensity. *Fishes* **8**(7) (2023). <https://doi.org/10.3390/fishes8070347>
10. Papini, O. *et al.*: Guidelines for the annotation of *Nephrops norvegicus* UWTV videos, (2025). <https://doi.org/10.5281/zenodo.14973160>
11. Melvin, É.: *Nephrops (Nephrops norvegicus) Burrow object detection simple training dataset from Irish Underwater TV surveys*. Zenodo, July 2024. <https://doi.org/10.5281/zenodo.13987957>.