Species distribution models (SDMs) of Common Redstart (*Phoenicurus phoenicurus*) – workflow and tutorial

Summary of the workflow

We use species distribution models (SDMs, hereafter called *Redstart model*) to address urban green space recommendations by predicting the suitable habitat of the Common Redstart (*Phoenicurus phoenicurus*) with a 2×2m grid cell resolution. A method and a tutorial are here presented to apply the Redstart model calibrated for the city of La Chaux-de-Fonds (Switzerland) and based on bird census data to other similar urbanized areas. The Common Redstart is used as a model species to develop planning recommendations for sustainable urban green spaces with tall indigenous trees as a main landscape element and conservation item. The aim of the Redstart model is to identify priority areas for conservation and promotion of the urban biodiversity. Consequently, the Redstart model is used under two scenarios: 1) the **current scenario** where the current predicting variables calculate a **conservation area**, i.e. an area that should be maintained in the current state and 2) the **promotion scenario** simulating an increase of the tree canopy density to calculate a **promotion area**, i.e., an area which can be improved to a suitable habitat for urban biodiversity by appropriate conservation measures.

The Redstart model is described in Droz et *al.* (2019)[1] and landcover categories follow the classification made in Droz et *al.* (2015)[2]. The workflow to build the predicting variables and to perform the Redstart model is described in this document. The R scripts are available on the https://github.com/Boris-Droz/Redstart-model-

Workflow

The *Figure 1* summarize the steps of the Redstart model workflow.

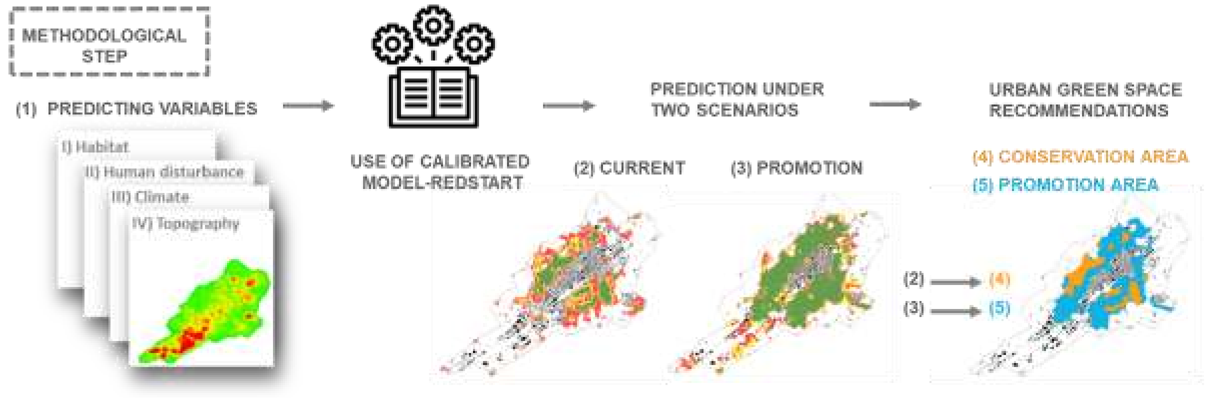


*Figure 1. Summary of the workflow. The methodological steps are 1) calculate the predicting variables; run the Redstart model under 2) the current predicting variables (**current scenario**) and under 3) a **promotion scenario** simulating an increase of the tree canopy density. The urban green space recommendations can be then applied on 4) conservation area and 5) promoting area calculated from 2) and 3) respectively.*

Input data

Seven spatial polygon or raster input data are needed to calculate the predicting variables for the **current scenario** and the **promotion scenario**. The link between the input data and the calculated predicting variables are presented in *Table 1* and are described in the sections below. The input data are recorded for the period between April – June which is the corresponding time of the bird census data. The model is based on data from Switzerland, but the Redstart model could be adapted for similar data of other areas with a resolution of 2×2 m grid cell or lower. When the resolution is finer than 2×2 m, our scripts downscale the resolution directly by averaging the values.

Table 1. Link between predicting variables and the input data to calculated them.

Rank	Pred. name	Input data
1	Trees canopy density	Lidar data
2	Impervious surface	Land cover
3	Human population density	Population count
4	Short-cut lawn	Land cover + Orthophoto with four spectral bands
5	High herbaceous vegetation	Land cover
6	Length of walls	Land cover
7	Traffic volume	Traffic count
8	Bare ground	Land cover + Orthophoto with four spectral bands
9	Solar radiation	Height Models (digital terrain model (DTM) and digital surface model (DSM))
	Conservation scenario	Urban green space recommendation

- **Height Models** were freely available at https://www.swisstopo.admin.ch/en/geodata/height.html. Product swissALTI3D and swissSURFACE3D were used as a digital terrain model (DTM) and a digital surface model (DSM) respectively. The 2016-2019 data set was used with a resolution of 10×10 cm grid cell, with elevation values in meter for the Swiss area. Data is available in a series of files with a swath dimension of 1 km^2 in geotiff or asc format.

- **Land cover** was freely available at https://www.geodienste.ch/ or provided by the canton geomatic agency. A polygon shape file was used for the year 2016 containing harmonized data with categories asphalted, buildings, green surfaces, water with subcategories[3].

- **Orthophotos with four spectral bands** were ordered from https://www.swisstopo.admin.ch/en/geodata/images/ortho/swissimage-rs.html. The 2016 Orthorectified image Swissimage RS swaths with four spectral bands (near infrared, red, green and blue) with color depth 16bits and with a resolution of 10×10 cm grid cell was used.

- **Point cloud Lidar data** were freely available at https://www.swisstopo.admin.ch/en/geodata/height.html. The data were classified following the American society for photogrammetry and remote sensing (ASPRS) nomenclature[5]. Files with a swath dimension of 1 km^2 with a resolution of 20x20 cm grid cell and an average of 15-20 pts/m^2 were used in their version 2016-2019.

- **Population count** were provided by the swiss federal statistical office (FSO) in two files: the data for population and households (STATPOP) statistics was ordered from https://www.bfs.admin.ch/bfs/en/home/statistics/population/surveys/statpop.html and the company statistic (STATENT) from https://www.bfs.admin.ch/bfs/fr/home/statistiques/industrie-services/enquetes/statent.assetdetail.11207837.html. Files contained geo-localized data point of the numbers of habitant and workers.

- **Traffic counts** were provided by the Federal Roads Office (FEDRO) for the year 2015. A polyline shape file contains the daily average traffic count (TJM) of the main roads (number of cars per day).

- **Urbanization land use plan** was provided by the canton geomatic agency. A polygon shape file was provided with categories of land use for each parcel. This layer is needed for the promotion scenario only.

Predicting variables

Table 2. List and description of the predicting variables

Rank	Pred. name	Code	Description	Value boundary	Unit
1	Trees canopy density	dtree	proportion cover of tree (high>5m)	0-100	%
2	Impervious surface	ddur	proportion cover of all house and asphalt (constructed surface)	0-100	%
3	Human population density	humcon	human concentration (habitant for each house, worker for each fabric and student for each school)	2.4-60.8	nb×m^{-2}
4	Short-cut lawn	dvegrase	proportion cover of short-cut lawn	0-100	%
5	High herbaceous vegetation	dveghaut	proportion cover of medium vegetation	0-100	%
6	Length of walls	lengthfas	sum of length wall for each bulding	0-0.8	m×m^{-2}
7	Traffic volume	mfroadw	proportion of car flux, considering surface between road also (wave perturbation)	2070-15250	car day^{-1} m^{-2}
8	Bare ground	dterrenu	proportion cover of bar ground	0-100	%
9	Solar radiation	fsumrad_2	averaged of the daily sum radiation (2m res scale)	0-17200	kJ×m^{-2}×day^{-1}

The names of the predicting variables (Pred. name) and codes (Code) are those used in the calibrated Redstart model and are consequently used in this document and in the R-scripts.

Calculation of predicting variable

Some changes have been made compared to the original analysis described in Droz et *al.* (2019)[1]. The calculation of some predicting variables has been optimized and automated thanks to new available datasets. We describe below the workflow adapted for the new predicting variables. The calculation of the other predicting variables remains the same and the description can be found in the supporting information in Droz et *al.* (2019)[1].

Trees canopy density: Point cloud Lidar data in LAS or LAZ format were analyzed with the lidR v3.1.1. package[4]. Classified point cloud data were restricted to Lidar class 3 with height equal or higher than 5 m and with NDVI between 0.5 to 1. A grid canopy was calculated with a spatial resolution of 0.5 m and a digital surface model algorithm using a 0.02 m radius circle. Then, a routine to detect trees using a 5m-radius circle window was used and segmentation was made using the Dalponte and Coomes (2016) algorithm (see detail in the lidR package documentation). The entire workflow was described in https://jean-romain.github.io/lidRbook/index.html.

Bare ground *and* ***Short-cut lawn***: Orthophotos with four spectral bands provided were used to distinguish bare-ground and short-cut lawn within the green area, i.e., permeable area cover with vegetation present in the land cover shape file (see above). First, we randomly selected 500 points inside the green area. Those points were manually classified as bare ground or short-cut lawn and the normalized difference vegetation index (NDVI) was calculated for those points:

$$\mathrm{NDVI} = (\mathrm{NIR} - \mathrm{RED}) / (\mathrm{NIR} + \mathrm{RED})$$

where NIR and RED are the spectral band measurements in the near-infrared and the red (visible) regions, respectively. We calibrated a binary generalized linear model (GLM) using our classified data point (bare ground vs short-cut lawns) as a function of the NDVI. Then, the calibrated NDVI-GLM was used to classify all pixel within the green area of the orthophoto. The mean NDVI were 0.078 and 0.417 for the bare ground and short-cut lawn respectively and the two classes were significatively different (*t*-test, *p*-value < 0.001). The NDVI-GLM performance was very good as evaluated by the area under the curve (AUC = 0.956) and the true skill statistic (TSS value = 0.82). False predictions were estimated to be 9% with a cross validation procedure. We used orthophotos produced and delivered from one flight in 2016 covering entire Switzerland. Therefore, the NDVI-GLM predict bare ground and short-cut lawn accurately independently of the region. However, we anticipate that bare ground and short-cut lawn detection may slightly change with other NDVI data sources, therefore we recommend to re-calibrate a NDVI-GLM to ensure a reliable detection.

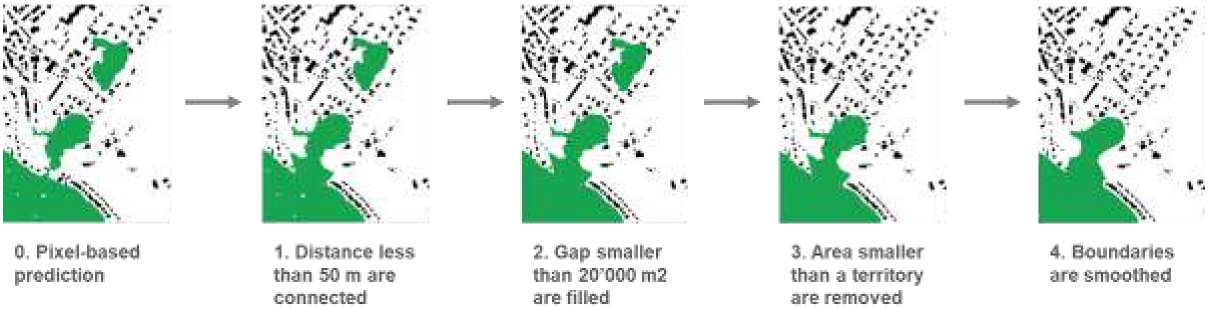
Solar radiation: We used the solar radiation model r.sun in RGrass v7.8.5 package because it is freely available. The radiation was calculated using the Linke turbidity factor, calculated elsewhere[5] and available in raster format and wgs84 projection in http://www.soda-pro.com/help/general-knowledge/linke-turbidity-factor. Measured values were used instead of predicted values for location as reported elsewhere[5].

The Redstart model to identify conservation and promotion area of the urban green space

We invite the reader to refer to our previous publication[1] for the detailed description of the Redstart model. We used species distribution models (SDMs) ensemble approach[6] combining four modeling techniques (two GLM, one GAM and one maxent) at very high-resolution scale (2×2 m grid cell) to predict the suitable habitat of the Common Redstart. One requirement to satisfy our prediction workflow is to have similar value boundaries for each given predicting variables as in Table 2. The closer the boundaries between the calibrate data set and the new predicted range are, the better the accuracy of the prediction will be. We considered a habitat as "optimal" if all the four modeling techniques predicted a suitable habitat, and as suboptimal if one to three models predicted it. To identify priority areas for the urban biodiversity in green space, the Redstart model is used under two scenarios: 1) the current scenario where the current predicting variables calculate a conservation area (i.e., optimal habitat) and 2) the promotion scenario simulating an increase of the tree canopy density to calculate a promotion area, i.e., an area which can be promoted to a suitable habitat for urban biodiversity under an appropriate management.

To convert the pixel-based prediction into conservation and promotion areas, we took in account additional conditions which we applied in the following order and the rules:

1. Areas with less than 50 m of distance are connected.
2. Gaps smaller than 20'000 m^2 within an area are filled.
3. Areas smaller than a territory (<31,400 m^2) are removed.
4. Area's boundaries are smoothed for a more straightforward display using the smooth function of the smoothr (version 0.2.1) package. Typical parameter as method and smoothness were set to ksmooth and 200 respectively (see package description for more detail).



0. Pixel-based prediction

1. Distance less than 50 m are connected

2. Gap smaller than 20'000 m2 are filled

3. Area smaller than a territory are removed

4. Boundaries are smoothed

Figure 2. Workflow to convert pixel-based prediction into conservation or promotion area.

Execution of the workflow

The production of predicting variables and the execution of the Redstart model is provided by a series of scripts. Its source code is primarily written in the R language version 4.0.3 exception made for the script Proj_ensemble_ Redstart.v5.0.R which runs under version 3.5.0. Some parts of the coding are supported by GRASS v.7 (https://grass.osgeo.org), Java Runtime Environment 64-bit 8.0-build-271 (http://www.filehippo.com/download_jre_64/) and Maxent version 3.3.3k (https://biodiversityinformatics.amnh.org/open_source/maxent/) which should be installed before starting. R library and automatic functions are compiled, updated and loaded during each script. Whenever it was possible, we implemented parallel calculations. The drawback of the parallel calculation was that the computer was sometimes too busy to perform other tasks. We recommend, if possible, to use a separate computer to run the scripts. Finally, we automatically removed the temporary raster at the end of each script to manage space on the computer. The workflow is supported by Windows 10 Pro platform but should work also under Linux with some slight modifications (e.g., on the folder path). The current workflow described here use the following R packages (in alphabetical order):

cpr version 0.2.3
dismo version 1.3-3
dplyr version 1.0.4
gam version 1.20
gbm version 2.1.8
geosphere version 1.5-10
grainchanger version 0.3.2
gstat version 2.0-6
lidR version 3.1.1
magrittr version 2.0.1
maptools version 1.0-2
units version 0.6-7
raster version 3.4-5
rgdal version 1.5-23
rgeos version 0.5-5
rgrass7 version 0.2-5
rJava version 0.9-13
rms version 6.1-1
RNetCDF version 2.4-2
sf version 0.9-7
smoothr version 0.2.1
sp version 1.4-5
spdep version 1.1-5
splitstackshape version 1.4.8
XML version 3.99-0.5

Tutorial

Whenever you start a new project to run the Redstart model in a new area, it is necessary to start this from a new *working directory* in copying the folder *Redstart-model* and rename it as you want. This directory will be your *working directory* containing already the necessary subfolder (input data, predicting variable, model projection) and R scripts to perform the entire workflow described above. It will also contain your output generated during the workflow.

Starting and performing a new project typically consists of

1. Copy the folder *Redstart-model* and rename it. This will be your working directory.
2. Install the add-ins mentioned in the *execution of the workflow* section above.

3. Prepare a shape file (shp) of the area where you want to run the Redstart model and put it on the *zone* subfolder.
4. Prepare or download your *input data*.
5. Calculate all predicting variable by running all script in the subfolder *script/pred* in chronological order.
6. Calculate the predictions by running the two scripts in the subfolder *script/proj* in chronological order.

In the following section, we describe all subfolder contained in your working directory.

Script folder – script and associate files

All scripts are organized similarly and contain a header with the script name and basic information (running R version, script version, coder, date). This is followed by a short description of what the script does, the important information about the script and the reference for codes taken from earlier works. Finally, a PARAMETER section provides the parameters that the user should or can change to correctly run the script. In the parameter section of each script, the working directory (*working_dir*), i.e, the folder path where is located the project folder is, should be set similarly for a given project. In each script, the code is provided below the declaration: ---- SCRIPT START HERE ---.

The folder script contains the following subfolder:

- **function**: Functions used in the scripts.
- **geographic_proj_model**: The calibrate Redstart model in R format.
- **linke_value**: Raster file of the Linke turbidity factor predicted elsewhere[5]: see *Solar radiation*. Raster data should be downloaded for April to June from the original site (http://www.soda-pro.com/help/general-knowledge/linke-turbidity-factor) and put in the *linke_value* subfolder.
- ndvi_model: The calibrated NDVI-GLM model in R format. See *Bare ground and short-cut lawn* under the section *Calculation of predicting values*.
- pred: Script to calculate the predicting variables.
- proj: Script to calculate the prediction of the suitable Common Redstart habitat and the conservation and promotion areas.

Input data folder – input

The *input* folder is an empty folder where you should put your input files to run the script from the folder *script/pred*. Each script has a description section for each input data needed. **Raster input** should be on one format supported by the raster package. Run `?raster:: writeFormats` R command, for the description of the file types supported. Preferred format is however geotif. **Shapefile input** should be provided on "ESRI shapefile" format. If not, conversion could be easily made with qgis (https://qgis.org) or others gis software. **Lidar data** should be provided preferentially in las format for fast reading. The R package lidR version 3.1.1 support las and laz format from .1 to 1.4.

Prior to start any script, make sure that **all input data are be provided with the same coordinate reference system** using epgs code (https://epsg.io/). If not, in the case of coordinate from the Swiss reference system, the swisstopo reframe calculator does the conversion (https://www.swisstopo.admin.ch/fr/cartes-donnees-en-ligne/calculation-services/reframe.html). In other cases, use qgis or others gis software to reproject coordinates.

Input data used to calculate some predicting variables are sometimes provided in multiple file vignettes for one area. In such a case, a folder containing all vignettes should be created in the folder *script/pred/.* The following folder names should be created for the corresponding input data and R scripts:

- *mnt* and *mns* (two separate folders): height models *1_merge_mnt_mnsv2.2.R.*
- *las*: lidar data *6_dtree_lidR_v1.2.R.*
- *rgbir*: orthophoto with four spectral bands *7_drase_dterrenu_v2.R.*

Install large data storage: Some of these input data require a lot of space on your hard drive (up to one Terabytes). Make sure you have enough remaining space on your hard drive for computing (at least 200 Gigabytes). If not, buy some solid state hard-drive to speed up the computing. To enable computing in a private computer containing limited amount of random-access memory (RAM), the script creates automatically a temporary subfolder in the *input* folder. The three following script are concerned: *1_merge_mnt_mnsv2.2.R, 6_dtree_lidR_v1.2.R* and *7_drase_dterrenu_v2.R*. The temporary subfolder are not automatically removed at the end of the script. Don't forget to check out and delete them to manage space on your hard drive.

Predicting variable folder – pred

The *pred* folder will contain all predicting variables generate from the scripts of the *script/pred* folder. The script 1 to 9 generate the predicting variables for the current scenario into the folder *pred/pred_cur.* Then, the script *11_pred_scen_v5.1.R* generate a subfolder per scenario increment. Computing the whole Redstart model with the script *1.proj_ensemble_Redstart.v5.2.R* for all steps might be very time consuming and not necessary. To determine the areas of conservation concern, computing the scenario *pred/pred_cur* and the *pred/pred_cons0.4* might be enough. Remove from the *pred* folder all unnecessary subfolder (scenarios) before running *1.proj_ensemble_Redstart.v5.2.R.*

Projection of the model folder – proj

The *proj* (projection) folder will contain all spatial predictions in geotiff format generate from the Redstart model and organized in one folder per scenario. Projection models are available for each modeling technique and replication as well as model weighted average raster called an ensemble (ENS).[6]

Performance of the workflow

Computing time

Computing time was evaluated in Window 10 Pro platform using a computer with an Intel i5 4300U processor (CPU @ 1.90GHz; RAM 11.7 GB 2 Core) and a solid-state hard drive speed of about 500 MBs. The computing time needed for each script is showed in

Table 3. As example, an area of 1 km^2 (25’000 cells) lasts 25x the time showed here.

Table 3. Computing time for each script.

	Script Names	time (s) / 10 000 cells
predicting var.	1 merge_mnt_mnsv2.2	5.0
	2 fsumrad_GRASSv2.2	114.8
	3 ddurv1.2	0.9
	4 ddveghautv1	0.5
	5 lengthfasv1	54.3
	6 dtree_lidR_v1.2	18.2
	7 drase_dterrenu_v2	23.4
	8 humconv1.2	6.6
	9 mfroadwv1	19.4
	10 made_zone_PA_v1	0.2
	11 pred_scen_v5.1	0.6
proj.	1 Proj_ensemble_Redstart.v5	6.1
	2 Zone_conserv_promov1	9.1

Redstart model performance and accuracy compared to the previous published model

A refreshed Redstart model has been recalibrated with new optimized and automated data sources. In the published version[1] the tree canopy density and bare ground input data were delimited by hand from orthophotos. Each of the four modeling methods calibrated with the new set of the nine predicting variables shows a good overall performance (high scores for AUC and TSS values), similar to the first Redstart model.[1] All modeling techniques exhibited AUC > 0.9 interpreted as excellent according to Araujo et al. (2005)[7] and TSS > 0.6. Maxent provided an excellent performance (mean AUC = 0.930; SD = 0.001 and mean TSS = 0.719; SD = 0.006) for the ten replications, followed closely together by GAM (AUC = 0.903, TSS = 0.658), GLM with model averaging (AUC = 0.905 and TSS = 0.671) and GLM with a stepwise selection (AUC = 0.906 and TSS = 0.673).

The new predicting variables, bare ground and tree canopy, present an overall better detection compared to the layer created manually. This is confirmed by some correct automatic detections not found in the layer created manually. This concern principally some very small, connected areas (less than 10 pixel) making altogether a relevant difference in the prediction. The better detection seems to improve the prediction resolution.

The current scenario prediction using the new calibrated Redstart model did not significatively differ from the previous Redstart model (*t*-test, p-value>0.1) where more than 75%-pixel prediction match the previous Redstart model. In detail, >90% of the conservation and promotion area predicted by three or fourth modeling technics are similar. Overall, 95% of the area predicted by the four techniques match the previous Redstart model.

The promotion scenario between the two Redstart models for La Chaux-de-Fonds did not differ significantly from the previous Redstart model (*t*-test, *p*-value>0.1). More than 95% of the area predicts by the four modeling technics are the same.

References

1. B. Droz, R. Arnoux, T. Bohnenstengel, J. Laesser, R. Spaar, R. Ayé and C. F. Randin, Moderately urbanized areas as a conservation opportunity for an endangered songbird, *Landscape Urban Plann.*, 2019, **181**, 1-9, https://doi.org/10.1016/j.landurbplan.2018.09.011.
2. B. Droz, R. Arnoux, E. Rey, T. Bohnenstengel and J. Laesser, Characterizing the habitat requirements of the Common Redstart *Phoenicurus phoenicurus* in moderately urbanized areas, *Ornis Fenn.*, 2015, **92**, 112-122, https://www.ornisfennica.org/pdf/latest/153Droz.pdf.
3. *WMS-MO Web Map Service avec les données de la mensuration officielle : Recommandations pour la réalisation - v1.5*, Conference des services cantonaux du cadastre, 2010.
4. J.-R. Roussel, D. Auty, N. C. Coops, P. Tompalski, T. R. H. Goodbody, A. S. Meador, J.-F. Bourdon, F. de Boissieu and A. Achim, lidR: An R package for analysis of Airborne Laser Scanning (ALS) data, *Remote Sens. Environ.*, 2020, **251**, 112061, https://doi.org/10.1016/j.rse.2020.112061.x.
5. J. Remund, L. Wald, M. Lefèvre, T. Ranchin and J. Page, H., Göteborg, Worldwide Linke turbidity, ISES Solar World Congress 2003 Sweden, 2003.
6. M. B. Araujo and M. New, Ensemble forecasting of species distributions, *Trends Ecol. Evol.*, 2007, **22**, 42-47, https://doi.org/10.1016/j.tree.2006.09.010.
7. M. B. Araujo, R. G. Pearson, W. Thuiller and M. Erhard, Validation of species-climate impact models under climate change, *Global Change Biol.*, 2005, **11**, 1504-1513, https://doi.org/10.1111/j.1365-2486.2005.001000.x.