Can a statistical regression be a valuable tool for the inversion of advanced IR sounders data?

A.B.Uspensky*, S.V.Romanov**, A.N.Trotsenko**

*Scientific and Research Center "Planeta",Roshydromet, Moscow, Russia
** Russian Research Center Kurchatov Institute, Moscow, Russia

Abstract

The paper summarizes the performance characteristics of linear statistical regression approach to the inversion of advanced IR-sounders data and retrieval of atmospheric temperature, humidity and ozone concentration profiles. The retrieval experiments with synthetic IASI data demonstrate the ability of proposed techniques to derive rather accurate sounding products at a reasonable cost.

Introduction

The new-generation IR-sounders on board future operational satellites (IASI/MetOp, CrIS/NPOESS, IRFS/Meteor) will provide high resolving power ($0.25 – 0.6$ cm^{-1}) spectral radiance measurements of continuous or quasi-continuous coverage from 3.7 (5.0) to 15.5 μm. Because of highly increased satellite data volume, large number of sought variables and data intrinsic properties the application of traditional inversion methodology ("physical" inversion, statistical regression) may cause significant problems. In particular, the major disatvantages of linear regression techniques application are as follows (Crone et al. 1996; Fleming et al. 1986; Rodgers 2000):

(i)   Neglecting the non-linearity of original inverse problem;

(ii)   Systematic biases in regression estimates caused by random errors in predictor variables;

(iii)   Gross error covariances of individual predictands induced by data multicollinearity (radiances in a number of channels are nearly linear dependent);

(iv)   Sampling effects.

In our study several refinements and modifications of standard regression technique have been developed and examined to alleviate negative effects (i), (iii), (iv) and to provide reasonably accurate retrieval of atmospheric temperature (T), water vapor (q) and ozone concentration (Q) profiles from IASI measurements (Chalon et al. 2001). The first approach is based on the Principal Component Analysis (PCA) technique. The known retrieval algorithms (Empirical Orthogonal Function or EOF regression) have been refined through the introduction of generalized empirical orthogonal functions both in measurement and state spaces.

The alternate (and sometimes complementary to PCA) approach to reduce the dimension of original inverse problem and thus to alleviate negative effects (iii), (iv) has also been investigated. It is based on the methods of selecting the most informative variables. Under many situations the critical issue for efficient regression performance is the choice of relevant predictors and predictands. For this reason, we have reexamined the use of physically valid predictors and predictands to account for non-linearity and interfering factors effect in original inverse problem. Along with this the problem of compilation the representative training datasets with minimal sample size is discussed.

A simulation study has been performed to demonstrate the capabilities of proposed techniques in the retrievals of T-, Q-, q- profiles from synthetic IASI measurements. The application of developed algorithms is shown to enable reasonable accuracy levels for T-, Q- profile retrievals close to those specified in IASI mission objectives (Chalon et al. 2001).

Methodology

We start with brief description of PCA and refined EOF regression algorithms, see for more details (Huang and Antonelly 2001; Uspensky et al. 2003b). Than the issues will be discussed relating to selection of the most informative variables as well as to the statement of relevant predictors in linear regression scheme. Finally, the technique to reduce the size of training datasets is described.

PCA technique

Each measured spectrum $\mathbf{R_j}$ subject to analysis (either compression or "inversion") is treated as n-dimensional random vector with known mean $\overline{\mathbf{R}}$ and covariance matrix S. The orthogonal basis in PCA is formed using normalized eigenvectors (e.v.) $\{\mathbf{u_\kappa}\}$ of covariance matrix S:

$$S\mathbf{u_\kappa} = \lambda_\kappa \mathbf{u_\kappa}, \kappa = 1, \ldots, n \quad \text{or} \quad SU = U\Lambda. \tag{1}$$

Here U is (n x n) matrix composed from n column-vectors $\mathbf{u_\kappa}$, and Λ is (n x n) diagonal matrix with eigenvalues $\lambda_1, \ldots, \lambda_n$ at principal diagonal.

If matrix S is not known , one can calculate the sample covariance matrix S_R, using spectral radiance measurements:

$$\mathbf{R_j} = \mathbf{R_{jt}} + \mathbf{e_j}, j = 1, \ldots, N. \tag{2}$$

Here N is a sample size, $\mathbf{R_{j,t}}$ are spectral radiance "true" values, and $\mathbf{e_j}$ is (nx1) vector of measurement errors with zero mean and known covariance matrix S_e.

Thus for each individual spectra $\mathbf{R_j}$ we have matrices S_R and S_e characterizing respectively the population variability and original data uncertainties. To account for the uncertainties in the data, the orthonormal basis $\{\mathbf{u_\kappa}\}$ from (1) is substituted by the new basis $\{\mathbf{v_\kappa}\}$ derived via solving the eigenproblem for so called "information operator" $G = S_R S_e^{-1}$ (Kozlov 1978; Thaker 1996):

$$GV = V\Lambda \text{ (or } S_e^{-1}\mathbf{v_k} = \lambda_k S_R^{-1}\mathbf{v_k}, k= 1, \ldots n). \tag{3}$$

To facilitate the solving of problem (3) with possibly asymmetric matrix G, one can introduce new variables $\mathbf{r_j} = S_e^{-1/2}\mathbf{R_j}$, where $S_e^{1/2}$ is the matrix S_e square root (Rodgers 2000). Solving the new eigenproblem for symmetric matrix $G' = S_r = S_e^{-1/2}S_R S_e^{-1/2}$

$$G'V = V \Lambda \text{ (или } G'\mathbf{v_k} = \lambda_k \mathbf{v_k}, k = 1, 2, \ldots, n) \tag{4}$$

results in the set of orthonormal e. v. $\{v_\kappa\}$, hereinafter referred to as generalized EOF's or GEOF's.

The compression or in other words the development of approximate model for the measurements (2) is performed using GEOF's as orthonormal basis. Each variation $\delta r_j = r_j - \overline{r}$ is fitted using least square estimation: $\delta r_{jtr} = V_{tr}\widetilde{\Theta}_j$, $\widetilde{\Theta}_j = V_{tr}^T \delta r_j$, where ($n_c$ x n) matrix V_{tr} is composed from $n_c < n$ leading e.v. and Θ_j is (n_cx1) vector of principal components (PC's) or metric-based PC's (Thaker 1996). The $\Theta_{\kappa j}$ values represent decorrelated and orthogonal linear combinations of the original variables with decreasing variances λ_k.

The approximation or reconstruction root-mean-square error (r.m.s.e.) is defined as

$$\rho^2 = n^{-1} \cdot E [(\delta r_j - \delta r_{j.tr})^T (\delta r_j - \delta r_{j.tr})] = n^{-1} \sum_{n_c+1}^{n}\lambda_k.$$

Evident constraint $\rho^2 \leq 1$ allows to define the minimum dimension n_c, that guarantees the required rmse level.

EOF and GEOF regression

The linearized problem of data (2) inversion and vector $\mathbf{x}$ retrieval is stated as follows:

$$\delta R = A\delta \mathbf{x} + \mathbf{e}, \tag{5}$$

where (n x m) matrix A is Jacobian matrix and $\delta \mathbf{x} = \mathbf{x} - \overline{\mathbf{x}}$, $\overline{\mathbf{x}}$ is a mean vector.

The special orthogonal basis is formed by e.v. of information operator $G = S_{o.x} \cdot S_x^{-1}$, where $S_{o.x}$ is covariance matrix for the ensemble $\{x_j\}$ and $S_x = M^{-1}$, $M = A^T S_e^{-1} A$ is Fisher information matrix.. Corresponding eigenproblem, analogous to (3), is written as

$$GW = W\Lambda \text{ or } Mw_k = \lambda_k S_{o.x} w_k, \quad k = 1,\dots,m. \tag{6}$$

The following modification of G regards to substitution of original normalized (noise scaled) spectra δr_j by compressed ones $\delta r_{j.tr}$. As a result, matrix M is replaced by matrix $M' = A^T S_e^{-1/2} V_{tr} V_{tr}^T S_e^{-1/2} A$ and matrix G is replaced by matrix $G' = S_{o.x} M'$.

Since matrices G, G' can be asymmetric, it is rational beforehand to transform original variables as follows:

$$r = S_e^{-1/2} R, \quad h = S_{o.x}^{-1/2} x.$$

The eigenproblem (6) becomes the problem with symmetric matrix $G'' = S_{o.x}^{1/2} M' S_{o.x}^{1/2}$:

$$G''W = W\Lambda. \tag{7}$$

Derivation of the eigenvalues $\{\lambda_a\}$ provides the specification of dimension $m_c < m$ for the GEOF basis W_{tr} consisting of $m_c = \max\{a : \lambda_a \geq 1\}$ leading e.v. $\{w_k\}$.

The decompositions $\delta h_{j.tr} = W_{tr}\Theta_{j.h}$ and $\delta r_{j.tr} = V_{tr}\Theta_{j.r}$ can be used in the GEOF regression algorithm:

$$\Theta_{jh} = C\Theta_{jr}, \quad \delta h_j = W_{tr}\Theta_{j.h} = W_{tr}CV_{tr}^T\delta r_j, \tag{8}$$

where C is ($m_c \times n_c$) matrix of regression coefficients (Smith and Woolf 1976; Uspensky et al. 2003b). The major advantage of (8) is that dimensions m_c, n_c are consistent to the informativity of original data (2) with respect to x.

Selection of the most informative variables

There exist at least two approaches to reduce the number of predictors: the transformation of original variables (like above PCA technique) and the selection of the most informative variables (channels), see e.g.(Aires et al. 2002; Rodgers 1996; Rabier et al. 2002; Uspensky et al. 2003a).The PCA methods may lead to the selection of a few linear combinations, which perfectly predict the variables of interest; but along with this the PC's depend upon all R_j components and their introduction doesn't provide real reduction in the number of measured quantities. The efficient reduction of the inverse problem dimension can be provided via application the methodology of the selection the most informative variables (channels). The recurrent algorithms of direct selection of informative variables (Rodgers 1996; Rabier et al. 2002) or exchange-type procedures of searching optimal channels (similar to technique applied in experimental design theory) may be utilized.

Starting from our experience, it seems reasonable to apply less formalized but more "physically-grounded" methodology. The proposed algorithm of optimal channels subset selection consequently analyses the information content (sensitivity of various channels to sought parameters), the altitude dependence of weighting function maximums, as well as level of the main interfering factors contribution to the measured radiance in each initially selected IASI channel (Trotsenko et al. 2003). The example of channel subsets selection (using described technique) is given below.

The choice of relevant variables in regression procedures

As mentioned in the Introduction, the specification of relevant predictor/ predictand variables helps under many situations to overcome (at least, partially) the non-linearity of original inversion problem and to account for the effect of interfering factors. Because listed features are mounting in case of Q-, and q- profile retrievals, the standard linear regression approach as well as EOF regression could become inefficient. To reduce the impact of surface temperature T_s and T- profiles variations (as key interfering factors) on the measured radiances the refined regression scheme should consider a priori

knowledge for these factors at the level of 0.5K and 1.5 K/km, respectively. Basing on these assumptions, the expression for the radiance in particular channel could be linearized respecting the reference state vector (X_0) as follows:

$$R = \mathbf{B}^T \mathbf{j}_0 + \mathbf{B}^T d\mathbf{j} , \qquad\qquad (9)$$

where R signifies the measured radiance correspondent to target state vector $\mathbf{x}$; $\mathbf{B}$ is a vector composed of the Planck functions for the T_s (the first element) and the T-profile correspondent to state $\mathbf{x}$; $\mathbf{j}_0$ is a vector containing full atmosphere transmittance as the first element and accordingly normalized transmittance derivatives respecting altitude coordinates correspondent to reference state vector $\mathbf{x}_0$, and $d\mathbf{j}$ is the deviation of vector φ respecting $\mathbf{j}_0$. In line with (9) the vector of predictors is composed of the values of $\Delta R = R - \mathbf{B}^T \mathbf{j}_0$ specified for each channel and expressed in units of equivalent brightness temperature. Namely vector $\mathbf{B}$ is composed on the basis of preceding retrievals for the T_s and the T-profile whereas vector $\mathbf{j}_0$ is produced (individually for each channel) by correspondent averaging of the pre-calculated transmittances and their derivatives over the learning sample. In turn the vector of predictands initially contains the values of $\ln(x_i / x_i^0)$, i = 0, 1,… L, where L is the number of atmosphere levels in the profile, and elements x_i denote the ozone (or water vapor) column integrated contents and relevant mixing ratios at i = 0 and i > 0, respectively. Note that just introduced predictor/predictand variables are, of course, not unique. In the next section the example will be given of q-profile regression retrieval with other predictor variables. Besides, the option to apply EOF decomposition for introduced predictor/ predictand variables is proved to be acceptable.

How to reduce the size of training samples?

The accuracy of regression estimates depends crucially on the quality of training samples. In order to decrease sampling effects the special approach may be applied based on the ideas of quasi-statistical modeling and design theory.

Suppose the goal is to generate the representative sample $\{\mathbf{R_j}\}$ using known first and second moments $\overline{\mathbf{R}}$ =0 and S. Standard approach consists of usage well known formulae

$R_j= V_{tr} \Lambda_{tr}^{1/2} V_{tr}^T \xi_j$, j=1,…,N,

where ξ_j is a random n-dimensional vector, Gaussian- distributed with zero mean and identity covariance matrix.

In order to reduce radically the sample size N it is proposed to build the new sample R*=$\{\mathbf{R_j}\}$, j=1,…,N*, N*=n_c+1<<N as follows:

$R^* = V_{tr} \Lambda_{tr}^{1/2} V_{tr}^T H$,

where H is N*xN* Hadamard matrix composed of +1 and -1; moreover N* should be multiple to 4, for example, N* =20 or N* =40.

Simulation study results

Proposed algorithms have been evaluated in the series of retrieval experiments with synthetic IASI data incorporated into two special global datasets. The first dataset (named RIE) that has been complied by Eumetsat, comprises of about 2000 collocated pairs of "observed" (synthetic)spectra $\mathbf{R_j}$ and respective state vectors $\mathbf{x_j}$. The original dataset has been subdivided into statistically homogeneous ensembles relating to various latitude zones and sounding time, namely, SAS (SubArctic Summer), SAW (SubArctic Winter), MLS (Mid Latitude Summer), TRP (Tropics).

Dataset II incorporates about 20000 state vectors $\mathbf{x_i}$ extracted from well- recognized data catalogue NOAA 88/89 and ECMWF 60-level sampled database as well as respective "pieces"of synthetic spectra $\mathbf{R_j}$. Beforehand the clear-sky atmospheric models have been identified within mentioned

ensembles, using the analysis of the temperature and water vapor spatial derivatives behavior. The resulted ensemble of clear sky atmospheric models incorporates about 10000 implementations. The values $\mathbf{R_j}$ have been calculated using specially developed fast radiative transfer model, see (Trotsenko et al. 2001).

The retrieval experiments have been performed separately on the base of RIE dataset (EOF and GEOF regressions) and dataset II (linear regression with selection of informative variables and statement of relevant predictors/predictands). Moreover, the results are discussed, for the most part not reported in our preceding publications (Uspensky et al. 2003 a, b; Trotsenko et al. 2003).

RIE dataset

Q-profile retrievals

The radiance spectra within R3 band (1000-1070 cm^{-1}) and within R1 band (650-770 cm^{-1}) were used as predictors for the retrieval of atmospheric ozone profiles Q(z).

Experiment #1 Standard EOF regression algorithm; the PC's are introduced both in measurement and state spaces; the original predictor variables are from R3 band.

Experiment #2 The same as experiment #1, but original predictor variables from R3 band are supplemented by 40 radiances measured in informative channels from RI band.

Experiment #3 GEOF regression using GEOF's both in measurement and state spaces; the original predictor variables are from R3 band (the same, as in experiment # 1).
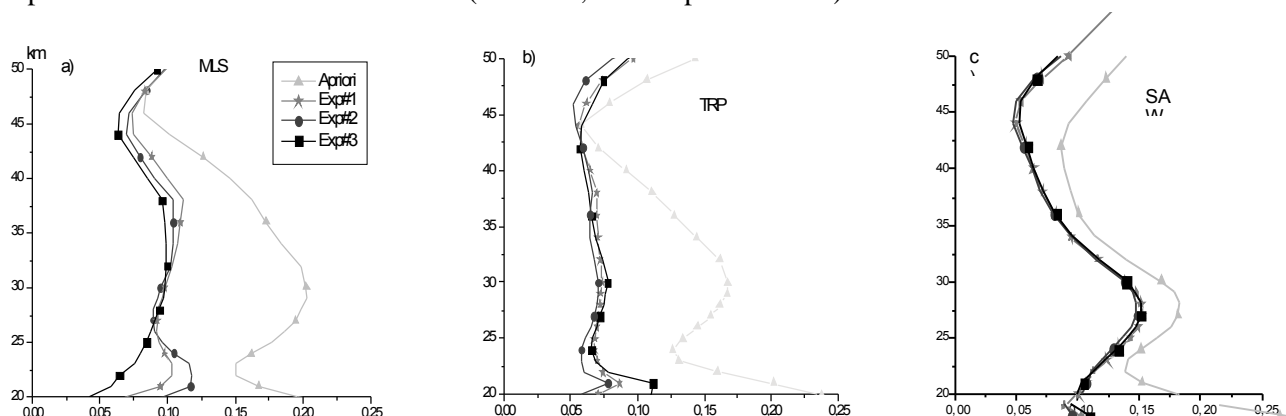

Fig.1 Ozone profile retrieval error statistics in Experiments #1-3

Fig. 1 (a, b, c) demonstrate error statistics for retrieval experiments #1-3 and control samples extracted from MLS (N=145), TRP (N=276), SAW (N=300) ensembles. Each fig contains three rmse profiles together with natural variability curve (square root of $S_{o.x}$ diagonal elements) that are plotted for the atmospheric layer within 20-50 km. It is not succeeded to achieve admissible level of retrieval accuracy (about 10 %) for SAW. Application of GEOF regression enables to improve the accuracy of ozone profile estimation only for MLS, while for the TRP ensemble the best results are provided by increasing the number of predictor variables (experiment # 2). Nevertheless in the majority of cases considered the use of developed technique enables to achieve the 10% accuracy levels of Q-profile retrievals.

T-profile retrievals

Retrieval experiments have been performed using the data from R1 band (481 IASI channels).

Experiment #1 Standard EOF regression; the PC's are introduced in both measurement and state spaces; the original predictor variables are from R1 band.

Experiment #2. The same as experiment #1, but the GEOF regression is applied.

As follows from the assessment of the layer mean r.m.s.e. values (defined via averaging over 3 layers: 0-10, 10-20, 20-30 km), the application of GEOF regression algorithm enables to improve slightly (<0.3K) the retrieval accuracy for the troposphere layers. Moreover, for all cases considered the developed algorithm ensures the retrieval of T-profiles with mean accuracy (mean r.m.s.e. about 1.3K within the layer 0-20 km.

q-profile retrievals

The retrieval technique, subject to testing, presents EOF regression algorithm with specially stated predictors/predictands. In contrast to above text dew point temperatures (DPT) were chosen as predictand variables, while the atmospheric brightness temperature (T_{AB}) were used as predictors. The T_{AB} values are derived as follows : $B_i(T_{AB}) = [R_i - B_i(T_s) \tau_{s,i}]/(1-\tau_{s,i})$, where $B_i(\ )$ and R_i are the Planck function and radiance in the i-th IASI channel; $\tau_{s,i}$ is the total atmospheric transmittance. The error statistics of the DPT regression estimation (T_{AB} in spectral region 1260-1360 cm^{-1} were replaced by 20 PC's and DPT profiles-by 15 PC's) is shown at fig.2a along with the a priori DPT variability for MLS atmosphere. The accuracy of the relative humidity (RH) profile retrieval is shown at fig.2b.
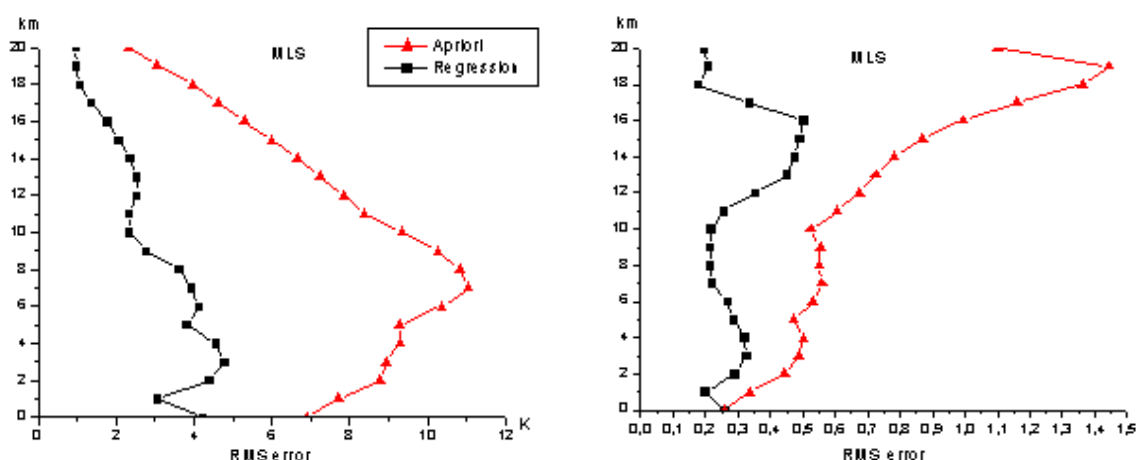

Fig. 2. The retrieval accuracy and a priori variability for DPT- and RH-profiles

As follows from fig.2, the accuracy of RH-profile retrievals is not high; therefore further refinements of retrieval technique are needed.

Verification experiments with dataset II

T- profile retrieval

The regression procedure for T-profile retrieval is based upon the use of IASI data in 55 selected channels (within band 652-743 cm^{-1}), see fig.3 below, illustrating the behavior of weighting functions in these channels. The use of optimal" channels subset permits to exclude the main interfering factors (the water vapor, the ozone, and the surface temperature). The retrieval tests show that the achieved rmse within 2 – 30 km altitude range is about or even better than 1 K. In turn the accuracy within the 0-1 km layer is not greater than 1.8 - 2 K. As illustration, the following fig.4 demonstrates the altitude dependence of the r.m.s.e. resulted from the retrieval test included 1213 independent implementations (70S-70N; 10W-45E).
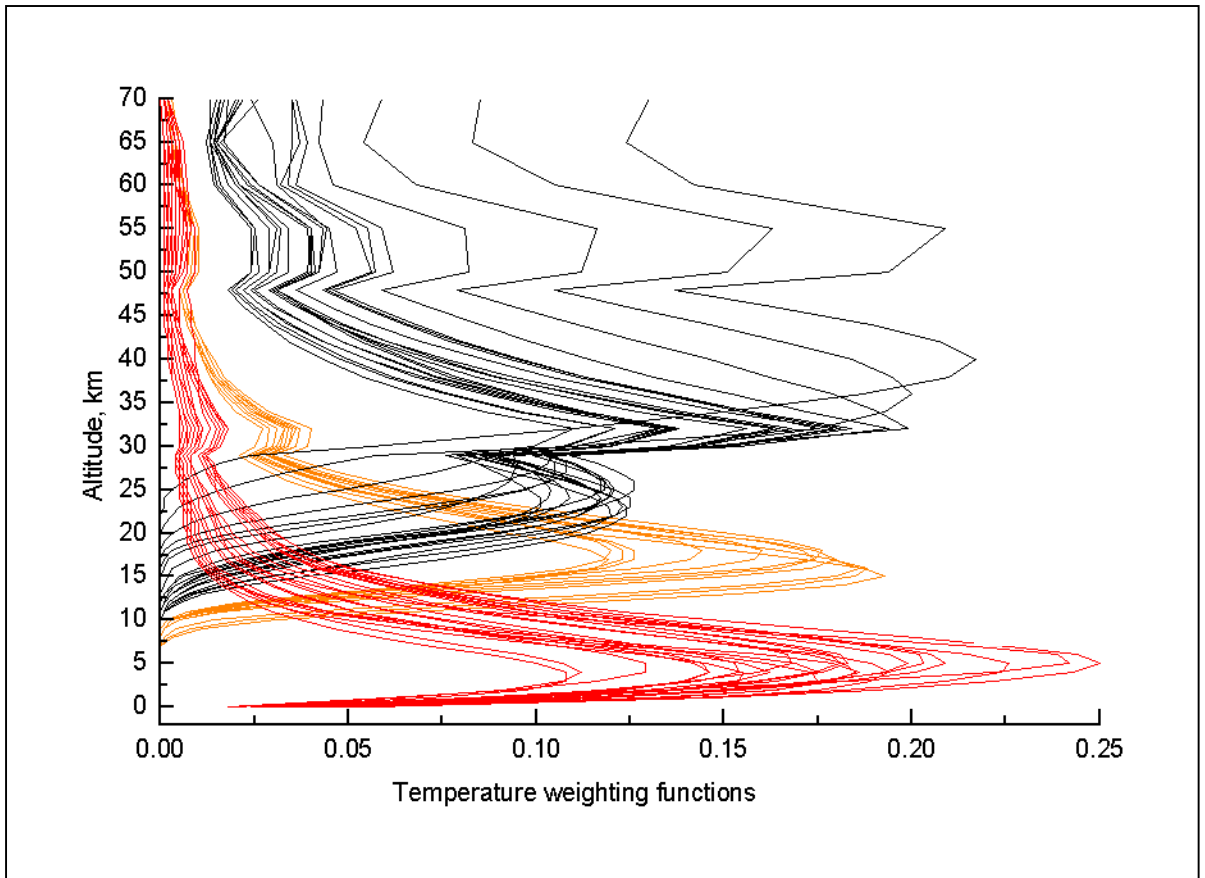
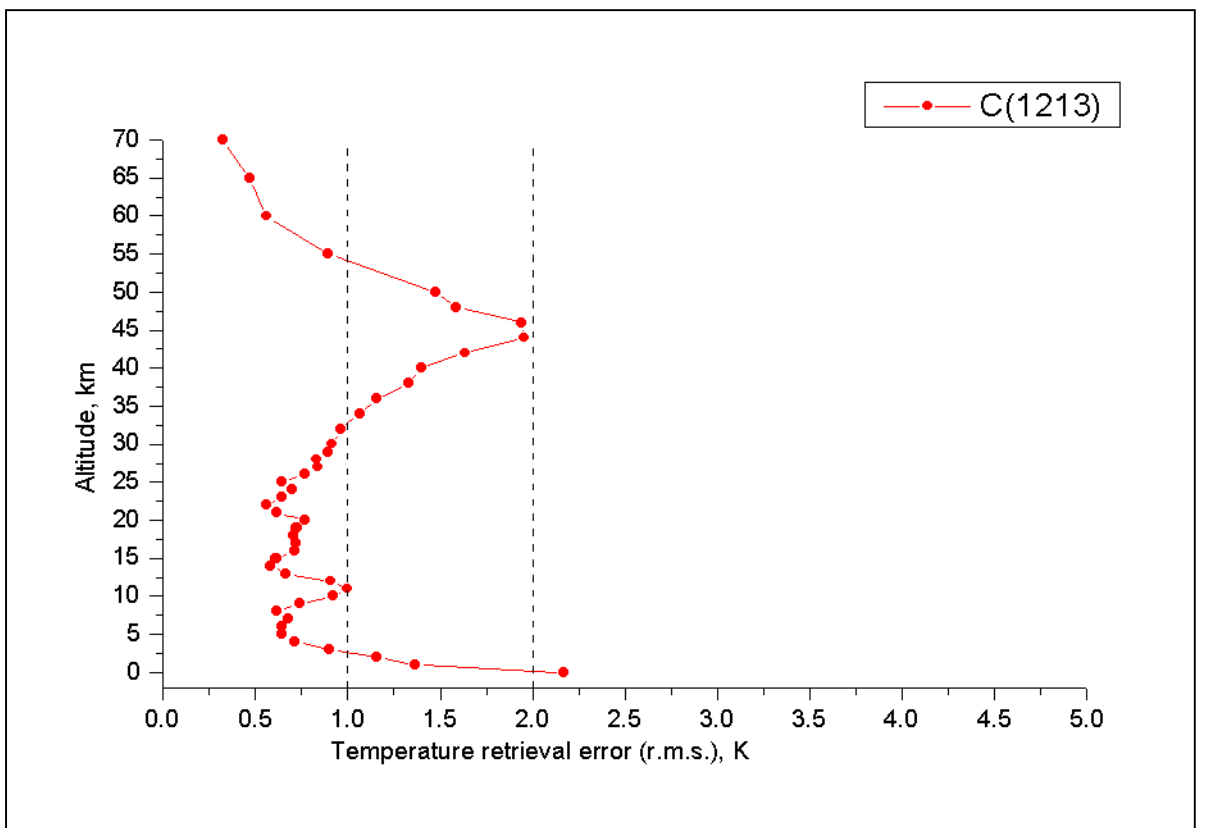Fig.3 Temperature weighting functions for 55 selected channels

Fig. 4. Error statistics for T-profile retrieval

Q-profile retrieval

The developed regression procedure utilizes the data in the dedicated channel subset as well as specially chosen predictor/predicatand variables (described above). The selection of the dedicated channel subset (universal for all atmospheric conditions) has resulted in the specification of 58 IASI channels centered withinR3 band. The composing of the training samples has been carried out according to prescribed sub-ranges of the ozone column amount. The altitude dependence of Q-profile retrieval accuracy (in terms of rmse) is shown at fig.5 (left panel) together with natural variability curves (right panel) for 6 samples relating to various atmosphere model ensembles. As it is illustrated on fig.5, the achieved retrieval accuracy within 20-50 km altitude range is about or even better than 10 %. Note that the comparison fig.5 with fig.1 manifests evident benefits of using dedicated subset of channels and relevant predictor/predictand variables.
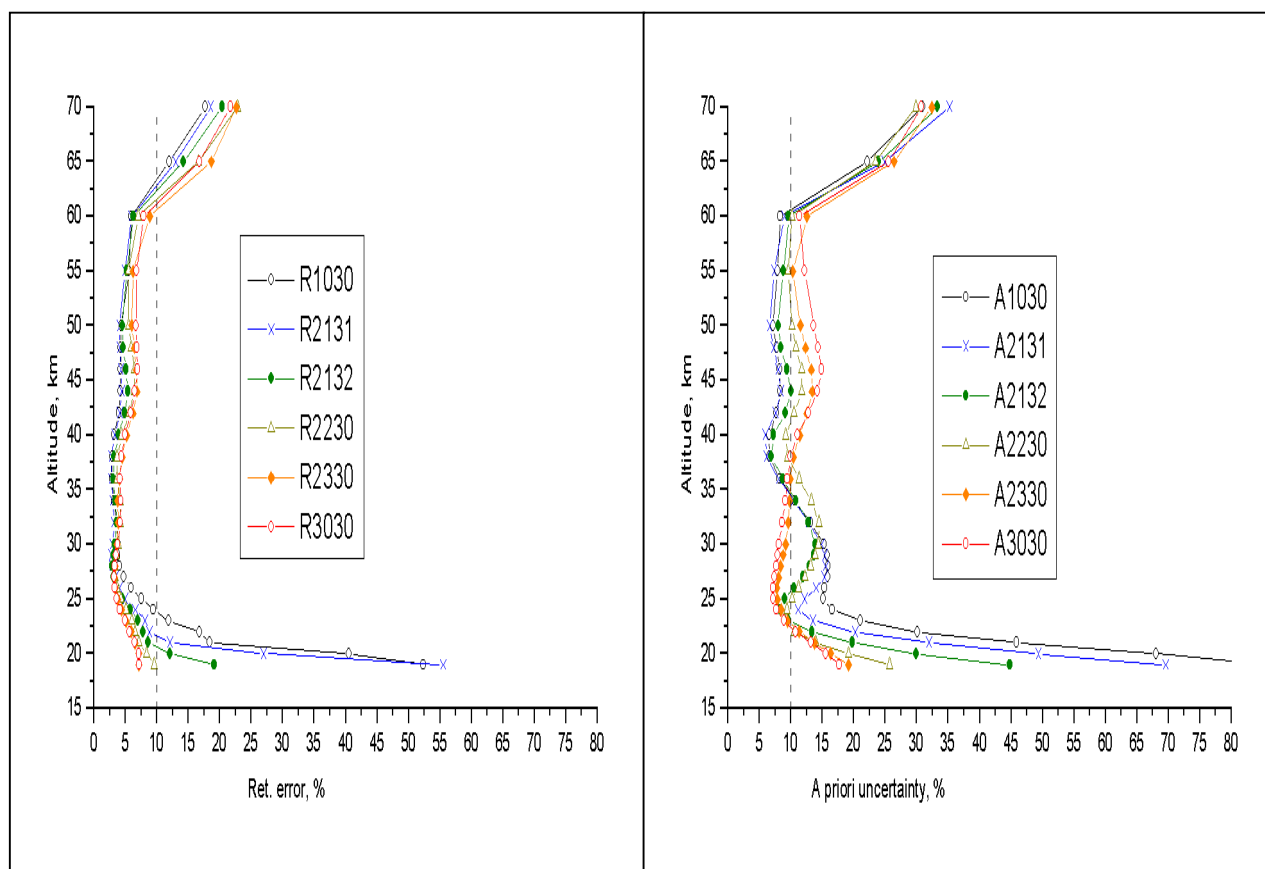

Fig.5 The retrieval accuracy and a priori variability for ozone profile

q –profile retrieval

The developed regression procedure is quite similar to Q-profile regression retrieval. Namely, it uses the data in the dedicated subset of 60 channels within 1206-1485 cm^{-1} band as well as predictor/predictands variables formally coinciding with those adopted for Q-profile retrieval scheme. At it has been shown by correspondent tests, the developed version provides rather good accuracy characteristics at least in the lower troposphere. This is illustrated by fig.6, where the r.m.s.e. altitude dependence (left panel) together with natural variability (right panel) are presented for 7 different samples. Comparison fig.6 with fig.2 demonstrates the advantages of proposed approach.
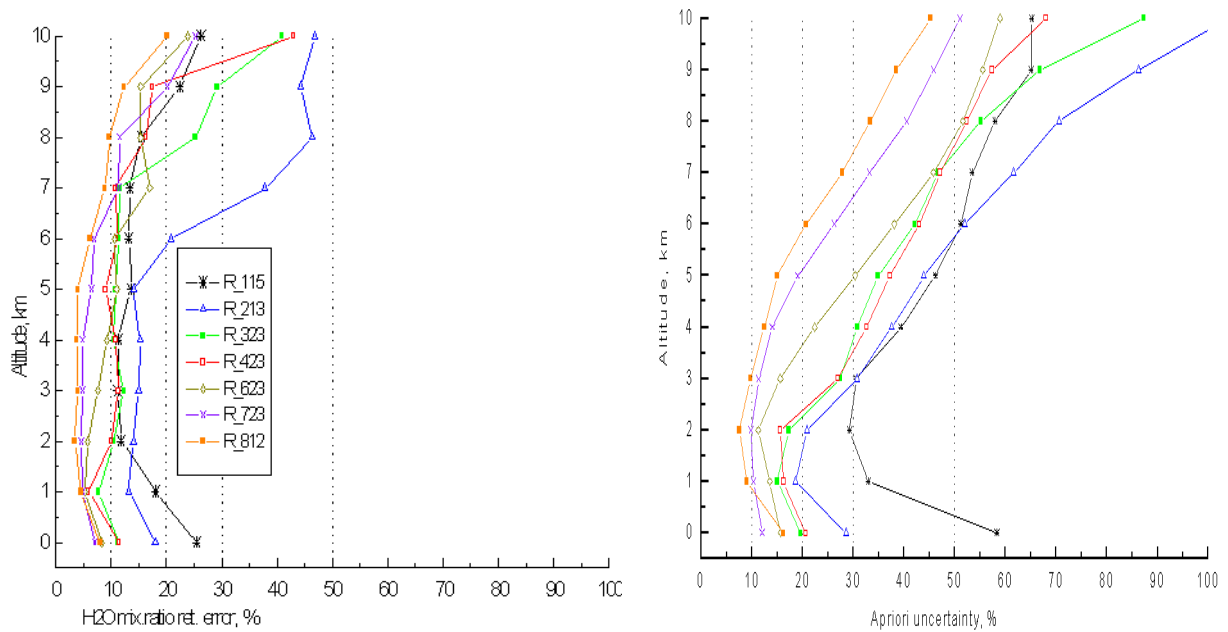
Fig.6 The retrieval accuracy and a priori variability for water vapor mixing ratio profile

Summary and conclusions

1.Statistical regression algorithms can definitely be a valuable tool for the inversion of advanced IR sounders data and retrieval of T-, q-, Q- profiles, if:

-to reduce significantly the number of original predictor variables as well as data multi-collinearity via introduction of PC's (EOF's, GEOF's) or selection of the most informative channels;

-to suppress (at least partially) the non-linearity effects through physically valid choice of predictors and predictands as well as through detailed classification of training samples;

-to reduce sampling effects through careful compiling of training samples, in particular, removing cases that are "cloudy"(especially for q- retrievals).

2. The simulation study with synthetic LASI spectra demonstrate the benefits of using the procedure of the most informative channels subset selection at the first stage of data inversion: the accuracies of output regression retrievals are better than those provided by "pure" EOF regression algorithms.

Acknowledgements

References

Aires F., Chedin A., Scott N.A., Rossow W.B. 2002. A regularized neural net approach for retrieval of atmospheric and surface temperatures with IASI instrument. J. Appl. Meteorol., 41, 144-159.

Chalon G., Cayla F., Diebel D. 2001. IASI. An advanced sounder for operational meteorology. CNES Rep., 9 p.

Crone L.J., MeMillin L.M., Crosby D.S. 1996. Constrained regression in satellite meteorology. J. Appl. Meteorol., 35, 2023-2-35.

Fleming H.E., Crosby D.S., Neuenderffer A.C. 1986. Correction of satellite temperature retrieval errors due to errors in atmospheric transmittances. J. Clim. Appl. Meteorol. v. 25, 6, 869-882.

Huang H.-L., Antonelly P. 2001. Application of principal component analysis to high-resolution infrared measurement compression and retrieval. Journ. Appl. Meteor., v.40, 365-388.

Kozlov V.P. 1978. Mathematical aspects of radiance measurements inversion. In "Abel inversion and its generalizations" Ed. Preobrazhensky N.G., Novosibirsk, Nauka, 68-95 (in Russian).

Rabier F.et al. 2002. Channel selection methods for IASI. Q. J. R. Meteorol. Soc., 128, 1011-1027.

Rodgers C.D. 1996. Information content and optimization of high spectral resolution measurements. In Optical Spectroscopy Technique and Instrumentation for Atmospheric and Space Research.II.SPIE, 23,136-147.

Rodgers C.D. 2000. Inverse methods for atmospheric sounding: theory and practice. World Sci. Publishing Co. 2000. 238 p.

Smith W.L., Woolf H.M. 1976. The use of eigenvectors of statistical covariance matrices for interpreting satellite sounding radiometer observations. J.Atm.Sci., v.33, 7, 1127-1140.

Thaker W.C. 1996. Metric-based principal components: data uncertainties. Tellus, 48A, 584-592.

Trotsenko A.N. et al. 2001. Efficient forward model to simulate IASI/METOP measurement and its validation against high-resolution field observations. In: IRS' 2000: Current Problems in Atmospheric Radiation, W.L.Smith, Ju.M.Timofeev (eds.), A.Deepak Publishing, Hampton,Virginia, USA, 401-404.

Trotsenko A.N. et al. 2003. Modular prototype processor for geophysical parameters retrieval from IASI: pilot version. The 2003 EUMETSAT Meteorol. Sat. Conf. Weimar, Germany, 29 Sept.-03 Oct. 2003 (In press).

Uspensky A.B., Romanov S.V., Trotsenko A.N. 2003a. Modeling of the atmospheric ozone vertical distribution remote measurements by the IR satellite sounders of high spectral resolution. J. Earth Research from space. Russian Acad. Sci., 1, 49-57 (in Russian).

Uspensky A.B., Romanov S.V., Trotsenko A.N. 2003b. Use of principal components technique for the analysis of advanced IR-sounders data. J. Earth Res. from space. Russian Acad. Sci., N 3, 26-33 (in Russian).

Proceedings
of the
Thirteenth International
TOVS Study Conference

INTERNATIONAL
ATOVS
WORKING GROUP
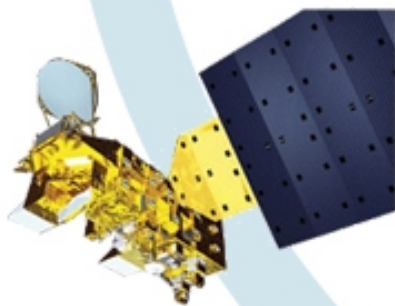
Sainte-Adèle, Québec, Canada
29 October – 4 November 2003