

Paper 155-31

Reading Compressed Text Files Using SAS® Software

Jaime Llano A., Independent Consultant, Bogotá, Colombia

ABSTRACT

External raw data files continuously needed at work are every day more and greater. Normally, those files come to you in any of the more popular compressed formats like zip, 7zip, gzip and rar. A common practice is to previously decompress files and next, read them with SAS®. Taking advantage of the PIPE device of the FILENAME statement and the aid of some tools free to use, you can directly read compressed data without extracting them first.

This paper shows different ways to read compressed data, considering several formats and software tools used to produce archives. Open code software will be used to access compressed data as well as a special dynamic library of the SAS Software.

All the SAS code used here has been tested with Windows XP SP2 and SAS 8.2. A special skill level from the reader is not required.

INTRODUCTION

When dealing with compressed files, there are two main software options to handle them: those with proprietary software license like Winzip, PKZip, Winrar etc., and others, free to use like 7Zip distributed under the GNU Library General Public License, and gzip a short for GNU zip, a GNU free software replacement for the Unix compress program. The tests you are going to see here will be using the last two ones, free to use: 7Zip and gzip.

The third option to use here is a dll installed in the sasexe folder of the core directory of SAS. It is an unsupported and undocumented option of the FILENAME statement, used in SAS to point to external files.

THE DATA

In order to allow an easy way to validate results, we have three text files named TEXT_FILE_01.txt, TEXT_FILE_02.txt and TEXT_FILE_03.txt. They all have the same register structure: CODE text 6, NAME text 25 and VALUE number 5. Additionally, each file has just three records.

Text files are compressed in archives with different formats. The table below shows the compressed files that will be used for testing, and the software used to produce them.

	Archive	Compressed With	Format
a	WinXP_Compressed_Files.zip	WinXP®	Zip
b	7Zip_Compressed_Files.7z	7zip	7z
c	7Zip_Compressed_Files.zip	7zip	Zip
d	gZip_Compressed_Files.z	Gzip	Gzip
e	pkzip_Compressed_Files.zip	pkzip®	zip

THE TOOLS

The software tools we will be using with SAS are:

SASzipam	dll installed with the SAS® System and used as option in the FILENAME statement
7Zip	open source file archiver that introduced the flexible 7z file format, but it also supports a number of other compressed and uncompressed archive file formats www.7-zip.org
gZip	does not archive files, it only compresses them, which is why it is often seen in conjunction with a separate archiving tool www.gzip.org

THE ENVIRONMENT

In order to assure that everything works well, let's put all you need in the same directory: the SAS programs we are going to test, the uncompressed text files, the archives (compressed files) and the tools 7za.exe and gzip.exe. This directory must be stated in the second line of our first program, and will be known as &MyDir (macrovariable).

TEST WITH AN UNCOMPRESSED FILE

The easy one first: reading the uncompressed version of TEXT_FILE_01.txt

```
Options nodate noxwait;
%Let MyDir = C:\xxxxxxxxxxxxxx;
x "cd &MyDir";
FileName Text01 "Text_File_01.txt";
Data DataSet00;
  Infile Text01 ;
  Input Code $Char6. Name $Char25. Value 5.;
  Put _All_;
run;
```

When submitted, the SAS log displays:

```
NOTE: The infile TEXT01 is:
      File Name=C:\SAS_JL_Papers\Text_File_01.txt,
      RECFM=V,LRECL=256
Code=Code01 Name=Name Code01 Text--File 01 Value=100 _ERROR_=0 _N_=1
Code=Code02 Name=Name Code02 Text--File 01 Value=200 _ERROR_=0 _N_=2
Code=Code03 Name=Name Code03 Text--File 01 Value=300 _ERROR_=0 _N_=3
NOTE: 3 records were read from the infile TEXT01.
      The minimum record length was 36.
      The maximum record length was 36.
NOTE: The data set WORK.DATASET00 has 3 observations and 3 variables.
NOTE: DATA statement used:
      real time          0.17 seconds
      cpu time           0.17 seconds
```

THE STRATEGY

The order to follow in these tests will be the following: 1) -Saszipam, 2)-7zip and 3)-gzip. For each one of them, test the reading of: a)-WinXP_Compressed_Files.zip, b)-Zip_Compressed_Files.7z, c)-7Zip_Compressed_Files.zip, d)-gZip_Compressed_Files.z and e)-pkzip_Compressed_Files.zip

SASZIPAM TESTS

1-A) WINXP_COMPRESSED_FILES.ZIP

```
FileName ComWinXP Saszipam "WinXP_Compressed_Files.zip";
Data DataSet01;
  Infile ComWinXP(Text_File_01.txt);
  Input Code $Char6. Name $Char25. Value 5.;
  Put _All_;
run;
```

When submitted, the SAS log displays:

```
NOTE: The infile library COMWINXP is:
      Stream=WinXP_Compressed_Files.zip
NOTE: The infile COMWINXP(Text_File_01.txt) is:
      File Name=Text_File_01.txt,
      Compressed Size=54,Uncompressed Size=114,
      Compression Level=-1,Clear Text=Yes
Code=Code01 Name=Name Code01 Text--File 01 Value=100 _ERROR_=0 _N_=1
Code=Code02 Name=Name Code02 Text--File 01 Value=200 _ERROR_=0 _N_=2
Code=Code03 Name=Name Code03 Text--File 01 Value=300 _ERROR_=0 _N_=3
NOTE: A total of 3 records were read from the infile library COMWINXP.
NOTE: 3 records were read from the infile COMWINXP(Text_File_01.txt).
NOTE: The data set WORK.DATASET01 has 3 observations and 3 variables.
NOTE: DATA statement used:
      real time          0.13 seconds
      cpu time           0.13 seconds
```

Everything is OK.

1-B) 7ZIP_COMPRESSED_FILES.7Z

```

FileName Com7Zipa Saszipam "7Zip_Compressed_Files.7z";
Data DataSet02;
  Infile Com7Zipa(Text_File_01.txt);
  Input Code $Char6. Name $Char25. Value 5.;
  Put _All_;
run;

```

When submitted, the SAS log displays:

```

ERROR: Cannot locate the end of the central directory in 7Zip_Compressed_Files.7z.
NOTE: The SAS System stopped processing this step because of errors.
WARNING: The data set WORK.DATASET02 may be incomplete.  When this step was
stopped there were 0 observations and 3 variables.
NOTE: DATA statement used:
      real time          0.29 seconds
      cpu time           0.22 seconds

```

Archives in 7z format cannot be read with saszipam

1-C) 7ZIP_COMPRESSED_FILES.ZIP

```

FileName Com7Zipb Saszipam "7Zip_Compressed_Files.zip";
Data DataSet03;
  Infile Com7Zipb(Text_File_01.txt);
  Input Code $Char6. Name $Char25. Value 5.;
  Put _All_;
run;

```

When submitted, the SAS log displays:

```

NOTE: The infile library COM7ZIPB is:
      Stream=7Zip_Compressed_Files.zip
NOTE: The infile COM7ZIPB(Text_File_01.txt) is:
      File Name=Text_File_01.txt,
      Compressed Size=112,Uncompressed Size=380,
      Compression Level=-1,Clear Text=No
ERROR: Invalid data length.
FATAL: Unrecoverable I/O error detected in the execution of the data step program.
Aborted during the EXECUTION phase.
NOTE: A total of 0 records were read from the infile library COM7ZIPB.
NOTE: 0 records were read from the infile COM7ZIPB(Text_File_01.txt).
NOTE: The SAS System stopped processing this step because of errors.
WARNING: The data set WORK.DATASET03 may be incomplete.  When this step was
stopped there were 0 observations and 3 variables.
NOTE: DATA statement used:
      real time          0.36 seconds
      cpu time           0.29 seconds

```

Archives in zip format produced with 7zip cannot be read with saszipam

1-D) GZIP_COMPRESSED_FILES.Z

```

Filename ComgZip Saszipam "gZip_Compressed_Files.z";
Data DataSet04;
  Infile ComgZip(Text_File_01.txt);
  Input Code $Char6. Name $Char25. Value 5.;
  Put _All_;
run;

```

When submitted, the SAS log displays:

```

ERROR: Cannot locate the end of the central directory in gzip_Compressed_Files.z.
NOTE: The SAS System stopped processing this step because of errors.
WARNING: The data set WORK.DATASET04 may be incomplete. When this step was
stopped there were 0 observations and 3
        variables.
NOTE: DATA statement used:
      real time          0.22 seconds
      cpu time           0.20 seconds

```

Archives in gzip format cannot be read with saszipam.

1-E) PKZIP_COMPRESSED_FILES.ZIP

```

Filename CompkZip Saszipam "pkzip_Compressed_Files.zip";
Data DataSet05;
  Infile CompkZip(Text_File_01.txt);
  Input Code $Char6. Name $Char25. Value 5.;
  Put _All_;
run;

```

When submitted, the SAS log displays:

```

NOTE: The infile library COMPKZIP is:
      Stream=pkzip_Compressed_Files.zip
NOTE: The infile COMPKZIP(Text_File_01.txt) is:
      File Name=Text_File_01.txt,
      Compressed Size=53,Uncompressed Size=114,
      Compression Level=-1,Clear Text=Yes
Code=Code01 Name=Name Code01 Text--File 01 Value=100 _ERROR_=0 _N_=1
Code=Code02 Name=Name Code02 Text--File 01 Value=200 _ERROR_=0 _N_=2
Code=Code03 Name=Name Code03 Text--File 01 Value=300 _ERROR_=0 _N_=3
NOTE: A total of 3 records were read from the infile library COMPKZIP.
NOTE: 3 records were read from the infile COMPKZIP(Text_File_01.txt).
NOTE: The data set WORK.DATASET05 has 3 observations and 3 variables.
NOTE: DATA statement used:
      real time          0.14 seconds
      cpu time           0.14 seconds

```

Everything is OK.

7ZIP TESTS

2-A) WINXP_COMPRESSED_FILES.ZIP ;

```

FileName ComWinXP Pipe '7za e "WinXP_Compressed_Files.zip"
                        "Text_File_01.txt" -y -so';
Data DataSet06;
  Infile ComWinXP;
  Input Code $Char6. Name $Char25. Value 5.;
  Put _All_;
run;

```

When submitted, the SAS log displays:

```

NOTE: The infile COMWINXP is:
      Unnamed Pipe Access Device,
      PROCESS=7za e "WinXP_Compressed_Files.zip"
      "Text_File_01.txt" -y -so,
      RECFM=V,LRECL=256
Code=Code01 Name=Name Code01 Text--File 01 Value=100 _ERROR_=0 _N_=1
Code=Code02 Name=Name Code02 Text--File 01 Value=200 _ERROR_=0 _N_=2
Code=Code03 Name=Name Code03 Text--File 01 Value=300 _ERROR_=0 _N_=3
Stderr output:

```

```

7-Zip (A) 4.23 Copyright (c) 1999-2005 Igor Pavlov 2005-06-29
Processing archive: WinXP_Compressed_Files.zip
Extracting Text_File_01.txt
Everything is Ok
NOTE: 3 records were read from the infile COMWINXP.
      The minimum record length was 36.
      The maximum record length was 36.
NOTE: The data set WORK.DATASET06 has 3 observations and 3 variables.
NOTE: DATA statement used:
      real time          0.49 seconds
      cpu time           0.18 seconds

```

Everything is OK.

2-B) ZZIP_COMPRESSED_FILES.7Z ;

```

FileName Com7zipa Pipe '7za e "7Zip_Compressed_Files.7z"
                          "Text_File_01.txt" -y -so';

Data DataSet07;
  Infile Com7Zipa;
  Input Code $Char6. Name $Char25. Value 5.;
  Put _All_;
run;

```

When submitted, the SAS log displays:

```

NOTE: The infile COM7ZIPA is:
      Unnamed Pipe Access Device,
      PROCESS=7za e "7Zip_Compressed_Files.7z"
      "Text_File_01.txt" -y -so,
      RECFM=V,LRECL=256
Code=Code01 Name=Name Code01 Text--File 01 Value=100 _ERROR_=0 _N_=1
Code=Code02 Name=Name Code02 Text--File 01 Value=200 _ERROR_=0 _N_=2
Code=Code03 Name=Name Code03 Text--File 01 Value=300 _ERROR_=0 _N_=3
Stderr output:
7-Zip (A) 4.23 Copyright (c) 1999-2005 Igor Pavlov 2005-06-29
Processing archive: 7zip_Compressed_Files.7z
Extracting Text_File_01.txt
Everything is Ok
NOTE: 3 records were read from the infile COM7ZIPA.
      The minimum record length was 36.
      The maximum record length was 36.
NOTE: The data set WORK.DATASET07 has 3 observations and 3 variables.
NOTE: DATA statement used:
      real time          0.33 seconds
      cpu time           0.17 seconds

```

Everything is OK.

2-C) ZZIP_COMPRESSED_FILES.ZIP

```

FileName Com7zipb Pipe '7za e "7Zip_Compressed_Files.zip"
                          "Text_File_01.txt" -y -so';

Data DataSet08;
  Infile Com7Zipb;
  Input Code $Char6. Name $Char25. Value 5.;
  Put _All_;
run;

```

When submitted, the SAS log displays:

```

NOTE: The infile COM7ZIPB is:
      Unnamed Pipe Access Device,
      PROCESS=7za e "7Zip_Compressed_Files.zip"
      "Text_File_01.txt" -y -so,
      RECFM=V,LRECL=256

```

```

Code=Code01 Name=Name Code01 Text--File 01 Value=100 _ERROR_=0 _N_=1
Code=Code02 Name=Name Code02 Text--File 01 Value=200 _ERROR_=0 _N_=2
Code=Code03 Name=Name Code03 Text--File 01 Value=300 _ERROR_=0 _N_=3
Stderr output:
7-Zip (A) 4.23 Copyright (c) 1999-2005 Igor Pavlov 2005-06-29
Processing archive: 7zip_Compressed_Files.zip
Extracting Text_File_01.txt
Everything is Ok

NOTE: 3 records were read from the infile COM7ZIPB.
      The minimum record length was 36.
      The maximum record length was 36.
NOTE: The data set WORK.DATASET08 has 3 observations and 3 variables.
NOTE: DATA statement used:
      real time          0.48 seconds
      cpu time           0.17 seconds

```

Everything is OK.

2-D) GZIP_COMPRESSED_FILES.Z

```

FileName Comgzip Pipe '7za e "gZip_Compressed_Files.z"
                          "Text_File_01.txt" -y -so';

Data DataSet09;
  Infile Comgzip;
  Input Code $Char6. Name $Char25. Value 5.;
  Put _All_;
run;

```

When submitted, the SAS log displays:

```

NOTE: The infile COMGZIP is:
      Unnamed Pipe Access Device,
      PROCESS=7za e "gZip_Compressed_Files.z"
      "Text_File_01.txt" -y -so,
      RECFM=V,LRECL=256
Code=Code01 Name=Name Code01 Text--File 01 Value=100 _ERROR_=0 _N_=1
Code=Code02 Name=Name Code02 Text--File 01 Value=200 _ERROR_=0 _N_=2
Code=Code03 Name=Name Code03 Text--File 01 Value=300 _ERROR_=0 _N_=3
Code=Code01 Name=Name Code01 Text--File 02 Value=100 _ERROR_=0 _N_=4
Code=Code02 Name=Name Code02 Text--File 02 Value=200 _ERROR_=0 _N_=5
Code=Code03 Name=Name Code03 Text--File 02 Value=300 _ERROR_=0 _N_=6
Code=Code01 Name=Name Code01 Text--File 03 Value=100 _ERROR_=0 _N_=7
Code=Code02 Name=Name Code02 Text--File 03 Value=200 _ERROR_=0 _N_=8
Code=Code03 Name=Name Code03 Text--File 03 Value=300 _ERROR_=0 _N_=9
Stderr output:
7-Zip (A) 4.23 Copyright (c) 1999-2005 Igor Pavlov 2005-06-29
Processing archive: gzip_Compressed_Files.z
Extracting Text_File_01.txt
Everything is Ok

NOTE: 9 records were read from the infile COMGZIP.
      The minimum record length was 36.
      The maximum record length was 36.
NOTE: The data set WORK.DATASET09 has 9 observations and 3 variables.
NOTE: DATA statement used:
      real time          0.35 seconds
      cpu time           0.19 seconds

```

It seems that the reading was OK but all the three text files of the "compressed" file were read. This must be that gzip compresses everything like a single file; does not archive files, it only compresses them.

2-E) PKZIP_COMPRESSED_FILES.ZIP -N+

```

FileName Compkzip Pipe '7za e "pkzip_Compressed_Files.zip"
                                "Text_File_01.txt" -y -so';

Data DataSet11;
  Infile CompkZip;
  Input Code $Char6. Name $Char25. Value 5.;
  Put _All_;
run;

```

When submitted, the SAS log displays:

```

NOTE: The infile COMPKZIP is:
      Unnamed Pipe Access Device,
      PROCESS=7za e "pkzip_Compressed_Files.zip"
      "Text_File_01.txt" -y -so,
      RECFM=V,LRECL=256
Code=Code01 Name=Name Code01 Text--File 01 Value=100 _ERROR_=0 _N_=1
Code=Code02 Name=Name Code02 Text--File 01 Value=200 _ERROR_=0 _N_=2
Code=Code03 Name=Name Code03 Text--File 01 Value=300 _ERROR_=0 _N_=3
Stderr output:
7-Zip (A) 4.23 Copyright (c) 1999-2005 Igor Pavlov 2005-06-29
Processing archive: pkzip_Compressed_Files.ZIP
Extracting Text_File_01.txt
Everything is Ok
NOTE: 3 records were read from the infile COMPKZIP.
      The minimum record length was 36.
      The maximum record length was 36.
NOTE: The data set WORK.DATASET11 has 3 observations and 3 variables.
NOTE: DATA statement used:
      real time          0.49 seconds
      cpu time           0.17 seconds

```

Everything is OK. There is a special tip with PKZIP compressed files. If the command that compressed the files used the option `-n+` file names will be stored in the long form. If not used, the same file names will be stored in the short DOS form (eight characters before the dot of the file name). If this is the case, we can still read it with SAS making the following adjustment:

2-EB) PKZIP_COMPRESSED_FILES.ZIP

```

FileName Compkzip Pipe '7za e "pkzip_Compressed_Files2.zip"
                                "Text_F~1.txt" -y -so';

Data DataSet10;
  Infile CompkZip;
  Input Code $Char6. Name $Char25. Value 5.;
  Put _All_;
run;

```

The name of the text file has only 8 characters before the dot. When submitted, the SAS log displays:

```

NOTE: The infile COMPKZIP is:
      Unnamed Pipe Access Device,
      PROCESS=7za e "pkzip_Compressed_Files2.zip"
      "Text_F~1.txt" -y -so,
      RECFM=V,LRECL=256
Code=Code01 Name=Name Code01 Text--File 01 Value=100 _ERROR_=0 _N_=1
Code=Code02 Name=Name Code02 Text--File 01 Value=200 _ERROR_=0 _N_=2
Code=Code03 Name=Name Code03 Text--File 01 Value=300 _ERROR_=0 _N_=3
Stderr output:
7-Zip (A) 4.23 Copyright (c) 1999-2005 Igor Pavlov 2005-06-29
Processing archive: pkzip_Compressed_Files2.zip
Extracting TEXT_F~1.TXT
Everything is Ok
NOTE: 3 records were read from the infile COMPKZIP.
      The minimum record length was 36.

```

```

    The maximum record length was 36.
NOTE: The data set WORK.DATASET10 has 3 observations and 3 variables.
NOTE: DATA statement used:
      real time           0.39 seconds
      cpu time            0.15 seconds

```

Everything is OK.

GZIP TESTS

3-A) WINXP_COMPRESSED_FILES.ZIP

```

Filename ComWinXP Pipe "gzip -cd WinXP_Compressed_Files.zip";
Data DataSet12;
  Infile ComWinXP;
  Input Code $Char6. Name $Char25. Value 5.;
  Put _All_;
run;

```

When submitted, the SAS log displays:

```

NOTE: The infile COMWINXP is:
      Unnamed Pipe Access Device,
      PROCESS=gzip -cd WinXP_Compressed_Files.zip,
      RECFM=V,LRECL=256
Code=Code01 Name=Name Code01 Text--File 03 Value=100 _ERROR_=0 _N_=1
Code=Code02 Name=Name Code02 Text--File 03 Value=200 _ERROR_=0 _N_=2
Code=Code03 Name=Name Code03 Text--File 03 Value=300 _ERROR_=0 _N_=3
Stderr output:
gzip: WinXP_Compressed_Files.zip has more than one entry--rest ignored
NOTE: 3 records were read from the infile COMWINXP.
      The minimum record length was 36.
      The maximum record length was 36.
NOTE: The data set WORK.DATASET12 has 3 observations and 3 variables.
NOTE: DATA statement used:
      real time           0.33 seconds
      cpu time            0.16 seconds

```

The last file of the archive was read. Gzip warns about the compressed file having more than one entry.

3-B) 7ZIP_COMPRESSED_FILES.7Z

```

Filename Com7zipa Pipe "gzip -cd 7Zip_Compressed_Files.7z";
Data DataSet13;
  Infile Com7Zipa;
  Input Code $Char6. Name $Char25. Value 5.;
  Put _All_;
run;

```

When submitted, the SAS log displays:

```

NOTE: The infile COM7ZIPA is:
      Unnamed Pipe Access Device,
      PROCESS=gzip -cd 7Zip_Compressed_Files.7z,
      RECFM=V,LRECL=256
Stderr output:
gzip: 7Zip_Compressed_Files.7z: not in gzip format
NOTE: 0 records were read from the infile COM7ZIPA.
NOTE: The data set WORK.DATASET13 has 0 observations and 3 variables.
NOTE: DATA statement used:
      real time           0.40 seconds
      cpu time            0.11 seconds

```

Archives in 7zip format cannot be read with gzip.

3-C) 7ZIP_COMPRESSED_FILES.ZIP

```

Filename Com7zipb Pipe "gzip -cd 7Zip_Compressed_Files.zip";
Data DataSet14;
  Infile Com7Zipb;
  Input Code $Char6. Name $Char25. Value 5.;
  Put _All_;
run;

```

When submitted, the SAS log displays:

```

NOTE: The infile COM7ZIPB is:
      Unnamed Pipe Access Device,
      PROCESS=gzip -cd 7Zip_Compressed_Files.zip,
      RECFM=V,LRECL=256
Code=Code01 Name=Name Code01 Text--File 01 Value=100 _ERROR_=0 _N_=1
Code=Code02 Name=Name Code02 Text--File 01 Value=200 _ERROR_=0 _N_=2
Code=Code03 Name=Name Code03 Text--File 01 Value=300 _ERROR_=0 _N_=3
Stderr output:
gzip: 7Zip_Compressed_Files.zip has more than one entry--rest ignored
NOTE: 3 records were read from the infile COM7ZIPB.
      The minimum record length was 36.
      The maximum record length was 36.
NOTE: The data set WORK.DATASET14 has 3 observations and 3 variables.
NOTE: DATA statement used:
      real time          0.45 seconds
      cpu time           0.15 seconds

```

The first file of the archive was read.

3-D) GZIP_COMPRESSED_FILES.Z

```

Filename Comgzip Pipe "gzip -cd gZip_Compressed_Files.z";
Data DataSet15;
  Infile ComgZip;
  Input Code $Char6. Name $Char25. Value 5.;
  Put _All_;
run;

```

When submitted, the SAS log displays:

```

NOTE: The infile COMGZIP is:
      Unnamed Pipe Access Device,
      PROCESS=gzip -cd gZip_Compressed_Files.z,
      RECFM=V,LRECL=256
Code=Code01 Name=Name Code01 Text--File 01 Value=100 _ERROR_=0 _N_=1
Code=Code02 Name=Name Code02 Text--File 01 Value=200 _ERROR_=0 _N_=2
Code=Code03 Name=Name Code03 Text--File 01 Value=300 _ERROR_=0 _N_=3
Code=Code01 Name=Name Code01 Text--File 02 Value=100 _ERROR_=0 _N_=4
Code=Code02 Name=Name Code02 Text--File 02 Value=200 _ERROR_=0 _N_=5
Code=Code03 Name=Name Code03 Text--File 02 Value=300 _ERROR_=0 _N_=6
Code=Code01 Name=Name Code01 Text--File 03 Value=100 _ERROR_=0 _N_=7
Code=Code02 Name=Name Code02 Text--File 03 Value=200 _ERROR_=0 _N_=8
Code=Code03 Name=Name Code03 Text--File 03 Value=300 _ERROR_=0 _N_=9
NOTE: 9 records were read from the infile COMGZIP.
      The minimum record length was 36.
      The maximum record length was 36.
NOTE: The data set WORK.DATASET15 has 9 observations and 3 variables.
NOTE: DATA statement used:
      real time          0.42 seconds
      cpu time           0.16 seconds

```

All files were read. It's ok. Remember gzip does not archive; just compress. Gzip must be used to compress a file at a time.

3-E) PKZIP_COMPRESSED_FILES.ZIP

```

Filename Comgzip Pipe "gzip -cd pkzip_Compressed_Files.zip";
Data DataSet17;
  Infile Comgzip;
  Input Code $Char6. Name $Char25. Value 5.;
  Put _All_;
run;

```

When submitted, the SAS log displays:

```

NOTE: The infile COMGZIP is:
      Unnamed Pipe Access Device,
      PROCESS=gzip -cd pkzip_Compressed_Files.zip,
      RECFM=V,LRECL=256
Code=Code01 Name=Name Code01 Text--File 01 Value=100 _ERROR_=0 _N_=1
Code=Code02 Name=Name Code02 Text--File 01 Value=200 _ERROR_=0 _N_=2
Code=Code03 Name=Name Code03 Text--File 01 Value=300 _ERROR_=0 _N_=3
Stderr output:
gzip: pkzip_Compressed_Files.zip has more than one entry--rest ignored
NOTE: 3 records were read from the infile COMGZIP.
      The minimum record length was 36.
      The maximum record length was 36.
NOTE: The data set WORK.DATASET17 has 3 observations and 3 variables.
NOTE: DATA statement used:
      real time          0.42 seconds
      cpu time           0.13 seconds

```

First file was read. Gzip warns about more than one entry.

CONCLUSIONS

Compressed text files can be read directly from SAS. The following table shows three alternatives:

		a	b	c	d	e
1	Saszipam	WinXP - zip	7Zip - 7z	7Zip - zip	gZip - z	pkZip - zip
2	7zip	OK	error	error	error	OK
3	gzip	OK¹	error	OK³	OK¹	OK²

Notes:

1. Compress only one file per gzip archive
2. In pkzip archives, file names depend on compression options
3. gZip reads first compressed file. Include only one file per gzip archive.

ACKNOWLEDGEMENTS

The author would like to express his appreciation to the following organizations for their support all these years using SAS software: SAS Institute Andean & Caribbean - SAS Institute Colombia S.A., SAS Institute Venezuela S.A. and SAS Institute Peru S.A.C., and especial thanks to Hugo Oliveros C.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Jaime Llano A. - Independent Consultant
 Bogotá, Colombia
 Phone: (571) 600-8656
 E-mail: jaimellanoa@yahoo.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.